BDM

BDM/VAS-MPC-96010

Folded Interconnection Network Development

Prepared by

Marc P. Chistensen

Michael W. Haney *

BDM Federal, Inc.

1996

Prepared for

Air Force Office of Scientific Research
&
Defence Advanced Research Projects Agency

* George Mason University

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
FINAL REPORT 15 Sep 92 - 14 Sep 96

4. TITLE AND SUBTITLE
FOLDED INTERCONNECTION NETWORKS DEVELOPMENT

5. FUNDING NUMBERS
8545/00
90NE198

6. AUTHOR(S)
Dr Haney

AFOSR-TR-97
0066

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
BMD International
4001 N. Fairfax Drive
Suite 750
Arlington, VA 22203

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NE
110 Duncan Avenue Suite B115
Bolling AFB DC 20332-8080

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-92-C-0062

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

The objective of the Folded Interconnection Network Development (FIND) program was to combine the rapidly emerging vertical cavity surface emitting laser (VCSEL) based "smart pixel" technology with a new free-space optical interconnection (FSOI) architecture which maximally exploits the ability of 3-D free-space optics to overcome the interconnection bottlenecks of multiprocessor systems. To focus the research, ultra-high throughput (-Tbit/sec) packet switching was selected as the key application area. The FIND program analytical results included a comparison of FSOI-based approaches with the traditional chip-multi-chip-module/printed-circuit-board metallic interconnection hierarchy. This analysis yielded fundamental scaling laws based on the geometrical consstraints associated with implementing high bisection bandwidth networks. The results proved that FSOI provides orders of magnitude advantage in size, weight, and power consumption for multiprocessor networks with bisection bandwidths greater than about 1 Tbit/sec. The experimental portions of the FIND program focused on the optomechanical issues related to the MCM-based retroreflective architecture. A one lens per chip design philosophy was adopted. Key optical elements of the package were evaluated with 1-D and 2-D arrays of VCSELs. An optical interconnection module was then designed and fabricated. The module achieved 10 micron resolution and registration accuracy across an entire 10 X 10 cm multichip substrate.

14

AGES

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	**18. SECURITY CLASSIFICATION OF THIS PAGE** UNCLASSIFIED	**19. SECURITY CLASSIFICATION OF ABSTRACT** UNCLASSIFIED	**20. LIMITATION OF ABSTRACT**

BDM/VAS-MPC-96010

Folded Interconnection Network Development

Prepared by

Marc P. Chistensen

Michael W. Haney *

BDM Federal, Inc.
4001 North Fairfax Drive, Suite 750
Arlington, VA 22203

November 1996

Contract F49620-92-C-0062

Prepared for

Air Force Office of Scientific Research
110 Duncan Avenue, Suite B115
Bolling AFB, DC 20332-0001
&
Defence Advanced Research Projects Agency Electronics Technology Office
3701 North Fairfax Drive
Arlington, VA 22203-1714

* Department of Electrical and Computer Engineering
George Mason University
Fairfax, VA 22030-4444

Table of Contents

1. Executive Summary

The ~~overall~~ objective of the Folded Interconnection Network Development (FIND) program was to combine the rapidly emerging vertical cavity surface emitting laser (VCSEL) based "smart pixel" technology with a new free-space optical interconnection (FSOI) architecture which maximally exploits the ability of 3-D free-space optics to overcome the interconnection bottlenecks of multiprocessor systems. To focus the research, ultra-high throughput (~Tbit/sec) packet switching was selected as the key application area, ~~although other multiprocessor DSP problems will benefit from the powerful approach that emerged from the program.~~

The FIND program ~~entailed extensive analytical and experimental efforts. The~~ analytical results included a comparison of FSOI-based approaches with the traditional chip/multi-chip-module/printed-circuit-board metallic interconnection hierarchy. This analysis yielded fundamental scaling laws based on the geometrical constraints associated with implementing high bisection bandwidth networks. The results proved that FSOI provides orders of magnitude advantage in size, weight, and power consumption for multiprocessor networks with bisection bandwidths greater than about 1 Tbit/sec. In related analysis, FSOI networks were represented as combinations of 6 basic topological transformations. The 3 key transformations developed under the FIND program include: 1. Using self-similar grids for the photonic I/O arrays to increase MCM packaging and optical efficiency, 2. Using a spatially interleaved I/O pattern for multistage networks to cluster I/O in a way that reduces control redundancy, and 3. Using a retroreflective architecture to permit all of the smart pixel resources to be located on a single multichip plane and simplify the optical alignment of the network. Implementation of this combination of transformations with 3-D optical elements provides clustering of I/O not possible with electronic implementations. The results proved that a dramatic reduction (~×50) in the redundancy of packet switch control and routing resources is possible with FSOI. This is significant because it is well accepted that, for large throughput switches, the control requirements can dominate the resource requirements. The topological considerations developed in the FIND program led to the invention of the Sliding Banyan

Network, which was shown in simulations to provide significant reductions in switching, control, and I/O resources over equivalent metallic-based approaches for realistic non-uniform traffic patterns. Variations on the control algorithm for the Sliding Banyan were evaluated using the industry standard *OPNET* network simulation tool to estimated smart pixel resource requirements. This was the first application of *OPNET* to a FSOI-based network architecture.

The experimental portions of the FIND program focused on the optomechanical issues related to the MCM-based retroreflective architecture. A one lens per chip design philosophy was adopted ~~to best match the multichip I/O clustering for the approach~~. Key optical elements of the ~~overall~~ package were ~~first~~ evaluated with 1-D and 2-D arrays of VCSELs ~~(supplied by Honeywell)~~. An ~~complete~~ optical interconnection module was then designed and fabricated ~~using catalog lenses and a mirr~~or. The module achieved 10 micron resolution and registration accuracy across an entire 10 × 10 cm ~~simulated 4 × 4~~ multichip substrate. A simplified alignment procedure was developed that is highly amenable to automation. The demonstration of the FIND module is the first optical interconnection module to demonstrate the required packagability across a large multichip photonic backplane.

The FIND program represents a significant advancement in the application of smart pixel and FSOI technologies to real-world problems. The analysis tools, design approach, and implementation techniques developed under this program provide a framework for the system level application of FSOI to significant multiprocessor problems in military and commercial arenas.

2. Program Overview

2.1 Introduction

This is the final technical report for the Folded Interconnection Network Development (FIND) program, covering the period September 1992 - September 1996. The FIND program combined the rapidly emerging Optoelectronic Integrated Circuit (OEIC) technology with a folded 3-D shuffle/exchange interconnection architecture for handling the communications needs in a parallel computer or telecommunications switch. The concept overcomes the bandwidth and density limitations of 2-D VLSI implementations by using 3-D free-space point-to-point links between smart pixel nodes containing optoelectronic I/O and electronic logic and memory circuitry.

Optical interconnection based systems have great potential for overcoming the interconnection bottlenecks inherent in multiprocessor systems. The potential application domain for free-space optical interconnections spans from multiprocessor switching systems to multiprocessor digital signal processing (DSP) systems. These probelems are characterized by large bisection bandwidths (BSBW) which severly limit performance and have great impact on the size weight and power of resulting systems. Packet switching, based on a shuffle-type interconnection topology, is a key emerging architectural approach to handling the high aggregate bandwidths of high performance switches. However, all-electronic switching approaches to packet switching have inherent performance limitations owing to the power, speed, and crosstalk constraints in VLSI technology. The smart pixel based FIND approach is a natural match to the performance and packaging needed for high performance switching.

The FIND concept is distinguished from other approaches in several ways. First, it is based on the use of imaging macro-optics that can be implemented with conventional refractive high performance optics that are used in an overlapped off-axis imaging architecture, thereby avoiding the long electrical paths needed for an electronic implementation. Second, it uses a new interleaving concept that allows the use of a single optical system and spatially multiplexed smart pixels to simultaneously implement all stages of a multistage interconnection network. This leads to some architectural advantages not possible in an all-electronic approach. Third, a novel array layout, developed during this reporting period and based on the self similarity property of fractal sets, permits the FIND concept to match well with Multi-chip Module (MCM) packaging schemes. Lastly, the FIND approach is ameanable to a novel retro-reflective implementation which is reduces volume and allows for a single MCM substrate, and leverages current automatable alignment techniques.

2.2 Objective

The overall objectives of the FIND program were:

- to evaluate the interconnection requirements of multiprocessor archtiectures and network switching fabrics in order to select an application or applications closely matched to a free-space optical interconnection system,
- to evaluate the performance of the free-space optical interconnection system based on reasonable projections of source/emitter array technologies,
- to define a free-space optical interconnection system based on the results of the evaluation and application study, and
- to demonstrate the feasability of a folded optical interconnection module.

2.3 Summary of Accomplishments

The FIND program acheived significant advances in the state-of-the-art of the application of free-space optical interconnections to multiprocessor problems. Under this progam the fundamental advanatages of free-space optics were identified and quatified to provide methodology for problem domains which reap maximum benefits from free-space optical interconnections (FSOI). These arguements are based on fundemental attributes of FSOI and have been experimentally validated during this program. These analyses provided the theoretical framework for the broad application of FSOI technology to systems. These theoretical benefits led to the invention of a new switching architecture, called the Sliding Banyan (SB). This architecture highlights the fundemental performance advantages of FSOI and points the way toward multi Terabit switching fabrics. The objectives of this program were acheived and exceeded through the following key accomplishements summarized below:

- The concept of self-similar grid patterns was introduced. Self-similar grid patterns were shown to provide a better match to the geometry of MCMs, thereby overcoming the packaging constraints associated with rectangular grids and providing a ×10 reduction in volume.

- The *Sliding Banyan* (SB) network was invented, patented and transfered to a commercialization organization (Capital Photonics Inc.). The concept was validated with 2D and 1D VCSEL arrays. Analytical validation proved dramatic reductions in the number of stages required to acheive a given blocking rate.
- The fundemental efficiency of the 3D SB architecture was extended to realistic area-of-interest traffic patterns. Simulations showed the SB approached the provable minimum switching resource requirement and reduced control resources by orders of magnitude.
- The SB optical interconnection module was experimentally validated. This prototype acheived 10 um resolution and registration accuracy across a 10 cm MCM like substrate. This was the first experimental demonstration of a retroreflective, multi-chip, single plane, FSOI concept.
- FSOI patterns were shown to be equivalent to topological transformations of graphs. This way of looking at FSOI provides a framework for exploiting the global interconnection benefits of FSOI. Six basic trnasformations were identified. The SB module combines all six of these to acheive its effieient packaging and performance.
- Algorithmic tradeoffs for control of the SB architecture were performed using the telecommunications model OPNET. This allowed a direct comparison of smart pixel complexity without implementation and fabrication of approaches.
- The fundemental geometric advantages of FSOI were quantified, for the first time. These advantages were based on projected capabilities of the electronic packaging hierarchy and paractical assumoptions about the abilites of FSOI. The results defined the application domain of FSOI to by the high BSBW domain of multiprocessor architectures. Any problem can be evaluated for its potential gain under and optical implementation by first defining its BSBW. These arguements show that a macro-optical single plane approach is the only one which scales well in size, weight, and power to the multi Terabit regime.

- An arbitrary interconnection approach based on macro-optics was invented. This interconnection fabric allows any multiprocessor architecture to derive advantages from FSOI, without regard to shuffle topology.

The details of these accomplishents are provided in Section 3.

3. Discussion of Results

3.1 Self-similar grid patterns in free-space shuffle/exchange networks

Optoelectronic free space links for shuffle/exchange networks were proposed to overcome the limitations of long metallic interconnections.[1] Versions based on 2-D arrays of processing elements (PEs) were suggested that more fully utilize the third dimension for higher density and efficient use of optical space bandwidth product.[2,3,4] Off-axis imaging techniques perform permutations such as the perfect shuffle[5] (PS) by optically interleaving equal sized sectors (such as quadrants) of the source array and overlaying the result onto an identical grid of detectors. In a Multistage Interconnection Network (MIN) the detected signals are then subjected to local exchange/bypass switching elements which operate on small groups of the array and route the signals to the next stage's source array. Various schemes offer trade-offs between the number of stages necessary for a given application and the complexity of the local switching elements. Examples of proposed optical networks include the 2-D separable PS,[2,3] the folded PS,[4] and higher order k-shuffles.[6] Figure 1 is an example of a 1-D PS, based on off-axis imaging and interleaving of two halves of the array. The figure is also a side view of the 2-D folded PS or 2-D separable shuffle, based on the off-axis imaging and interleaving of the four quadrants of the array. The optical efficiency for those optical sources located at the outer regions of a quadrant can be increased by adding a lens over each quadrant, as shown in the figure, to direct the cone of light from each pixel's source to the center of the associated off-axis imaging lens.

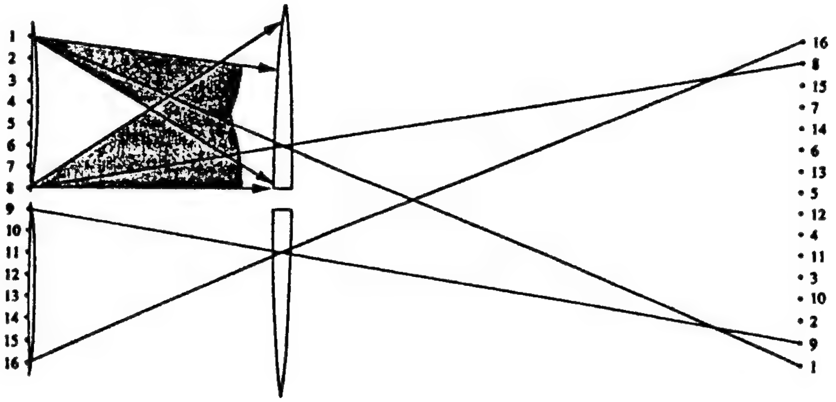


Figure 1. Side view of 2-D separable shuffle or folded perfect shuffle. The auxiliary lenses over the input plane improve optical efficiency by centering each pixel's beam on its imaging lenses.

The density of Optoelectronic Input/Output (OE I/O) circuitry on an OE Integrated Circuit (OEIC) chip is constrained by the chip area and heat dissipation requirements for the circuitry at each pixel. The off-axis imaging shuffle interconnects of the type depicted in Figure 1, however, dictate that each pixel must be positioned at a node of a rectangular grid array. Ideally, the circuit designer would be able to position the circuitry across the entire input plane, with the OE I/O on a rectangular grid, with a density limited only by thermal and other logic design considerations. In this case the maximum pixel density would be $1/p^2_{min}$, where p_{min} is the minimum allowed spacing between pixels, as determined by circuit constraints.

For very large arrays (i.e., 32×32 or greater) in which the smart pixels consist of OE I/O and several hundred or more electronic logic gates, we must assume that the pixel array will be distributed across several OEICs that are carefully aligned in the PE plane for optical interconnection. Even dense Multi-chip Module (MCM) chip packaging technology may require some separation between chips to allow for heat dissipation and metallic connections to the substrate's conducting paths. Other practical considerations which limit the density of chips stem from cost issues driven by the need for ease of design, manufacture, test, and repair. For an average high performance application, MCMs achieve about 40% area efficiency, defined as the ratio of active chip area to package area. Furthermore, this area efficiency is projected not to increase very much in the foreseeable future due to the above practical considerations.[7] If the pixels reside on a square grid, then the density of PEs is upper bounded by $1/d^2$, where d is the minimum spacing between chips on the MCM. For example, if the chip spacing is limited to d=.5 cm, and pmin is .1 cm, then the overall packing efficiency is limited to $(.1/.5)^2 = .04$ of the desired capability.

To overcome the density limitations we must array the smart pixels such that they are bunched together to accommodate a MCM die pattern for the OEICs, yet maintain the properly interleaved pattern after the shuffling optics. The proposed solution is based on the self-similarity of certain fractal sets.[8] The 1-D N pixel PS is considered first, then extended to 2-D PSs with N^2 element arrays, and finally generalized to higher order shuffles. Beginning with a line segment of length D, remove a central portion to leave two equal line segments of length ηD. From each of these two segments remove a percentage $1-2\eta$, and continue this process, successively removing the central $1-2\eta$ portion from each remaining segment. If the process is continued ad infinitum, the result is a fractal set. This type of set is self-similar, that is it is invariant to scale changes of $1/\eta$. To use such a set to form pixel grid patterns for shuffle exchange networks, the formation of the set is terminated at a point where the number of line segments is equal to the number of pixels to be arrayed in one dimension. A pixel location is assigned to the middle of each line sub-segment. Such an assignment is illustrated in Figure 2.

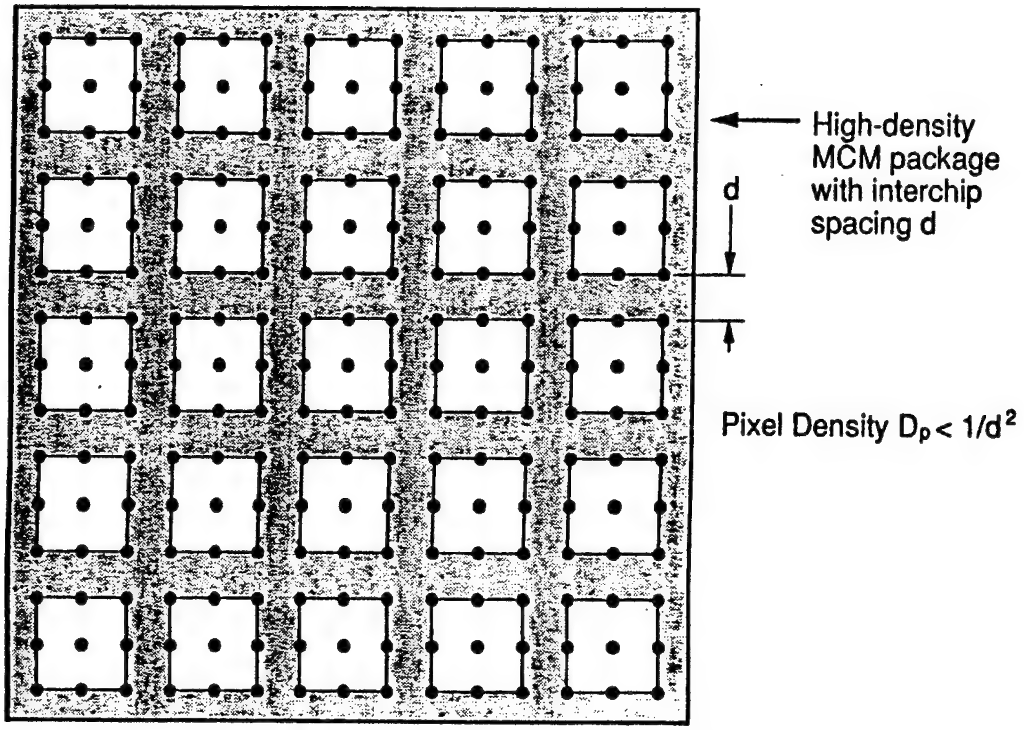


Figure 2. Illustration of pixel OE I/O placement based on a self-similar grid pattern.

Magnification of the array by $1/\eta$ and ignoring one side of it yields an identical array with every other element missing. If the other half of the array is also magnified and properly positioned over the first, then an identical array to the original is obtained, with the elements ordered in a PS of the original array, just as in the regularly spaced array of Figure 1. A geometrical analysis of the placement of each pixel of an N element array, with self-similarity parameter η, yields:

$$x_{pix} = D(\frac{1-\eta}{2}(\pm\eta \pm \eta^2 \pm \cdots \pm \eta^{n-1}), \quad 1)$$

where D is the width of the square PE array plane, and n=log$_2$N. Note that with η=.5 the pattern degenerates to a regular, equally spaced array, with spacing D/N. Figure 3 illustrates the side view lens placement necessary to achieve the self-similar array PS interleave pattern. This technique will have a higher light efficiency, owing to the bunching of OE I/O, and may avoid the need for the auxiliary lenses at the input plane shown in Figure 1. The pair of off-axis imaging lenses that perform the interleaving is placed to provide a magnification of 1/η. From geometrical considerations, their centers are located at positions off of the central axis given by:

$$X_{len} = \pm D\left[\frac{(1-\eta)(1-\eta^n)}{2(1+\eta)}\right]. \quad 2)$$

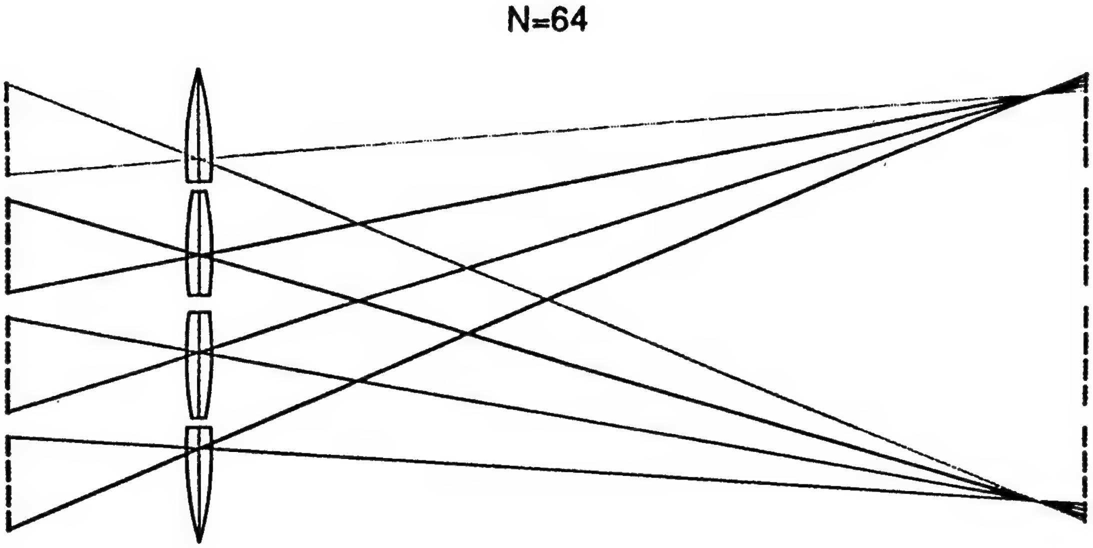


Figure 3. Side view of improved perfect shuffle scheme based on self-similar grid pattern with h = 1/3.

Note that this scheme naturally places paired pixels in close proximity to each other. This may be advantageous because each pair of signals will be input to the same local exchange/bypass switch. To obtain a 2-D array of N^2 smart pixels with this approach, we simply do the same thing in the orthogonal direction. Figure 4 depicts a

MCM substrate and chip array with a self-similar grid pattern ($\eta=1/3$) for the OE I/O superimposed. In this case the pixel patterns are identical at each chip site except for the 9 chips along the axes, which could be used for all-electronic ICs. The details of the design will be determined by the network's performance requirements, the OEIC capabilities, and the need for compatibility with MCM design techniques.

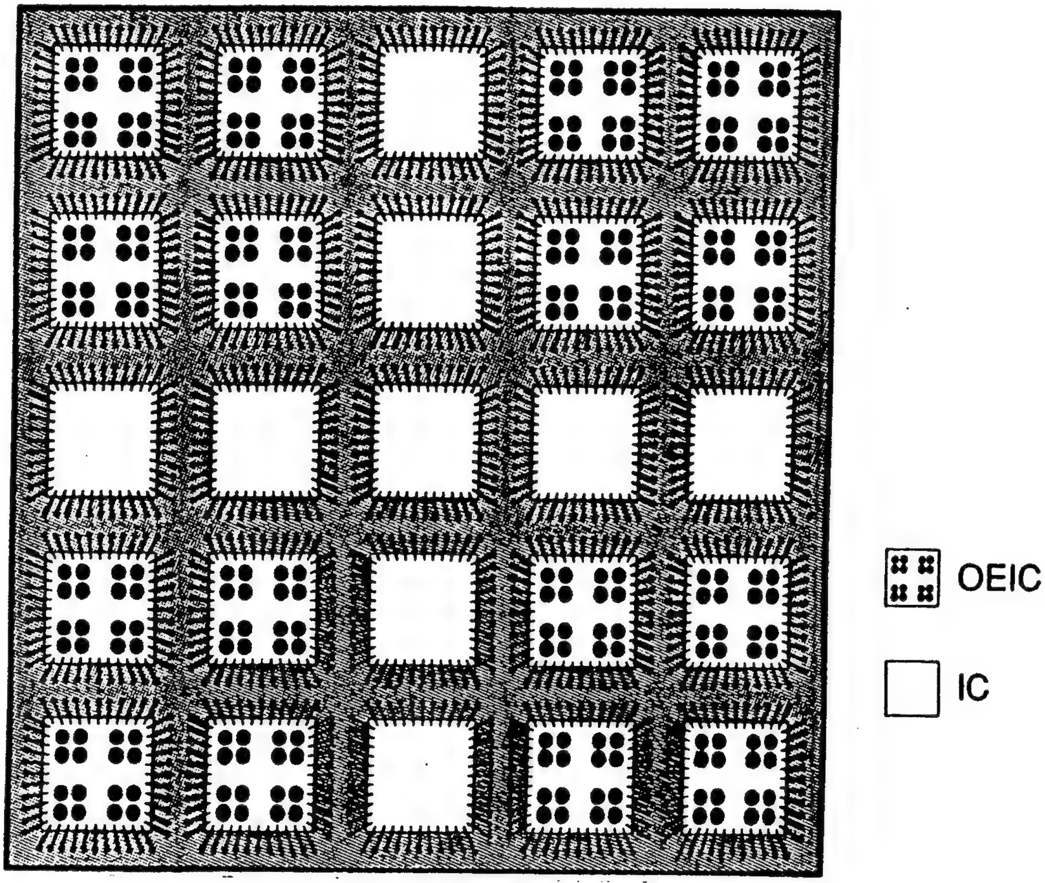


Figure 4. Example MCM chip layout with self-similar OE I/O grid pattern, with h = 1/3, for 256 smart pixels. In this case, the ICs along the axes of the MCM substrate contain no OE I/O.

The generalization of this approach to higher order shuffles is straight forward. In 1-D, whereas the PS follows from dividing and interleaving 2 sectors of the array, a k-shuffle follows from interleaving k equal sized sectors of the array. The 2-D optical implementation of a 2-D separable k-shuffle therefore requires k×k appropriately positioned off-axis imaging lenses. Following the above procedure for the PS (2-shuffle), we first consider a 1-D array. The unit line segment is first divided into k equally sized and equally spaced segments, each of which is further similarly divided, and so on. For a k-shuffle, the self similarity parameter, η, must be less than or equal to $1/k$. As before we extend the approach to 2-D arrays by using the same pixel placement scheme for the orthogonal dimension. The bunching of OE I/O in the k-shuffle provides improved optical efficiency in the same manner as for the PS described previously.

To estimate the improvement in pixel density provided by the self-similar grid approach, we first make the following assumptions. For the k-shuffle of N^2 PEs (where N is assumed to be a power of k) we assume there are M^2 identical OEIC chips with P^2 PEs on each chip (M and P are also assumed to be a power of k). We assume that optimum values for η and the square OEIC chip width are first calculated such that the closest pixels on an OEIC chip of the array are separated by pmin, and the closest OEIC chips are separated by d (typically with $d \gg p_{min}$). Then an estimate of the upper bound on the smart pixel packing density is given by the percentage of substrate area covered by pixels separated by p_{min} divided by p^2_{min}. From geometrical considerations, this is found to be:

$$P = \left[N \frac{\eta^{\gamma-1}}{kp_{\min}} \right]^2 , \qquad 3)$$

where $\gamma = \log_k N$. As a numerical example, consider a 2-D PS with an N^2=1024 pixel array, distributed across a 4×4 OEIC chip array (M=4), with 8×8 pixels on each chip (P=8). For d=.5 cm and pmin=.1cm, the optimum self similar parameter and chip size are derived to be approximately η=.413 and dchip=1.17cm, respectively. The area efficiency, from Equation 3, is calculated to be P=21.6 pixels/cm^2. When compared with the pixel density achievable with a regular square array, $1/d^2$=4 pixels/cm^2, we see an improvement of more than a factor of 5 in OEIC plane area to achieve the same array size. For a given speed of imaging optics, the optical volume required for this example is more than a factor of 10 smaller than the equivalent square grid array femdash a sizable savings.

The rapid advances being made in OE technology, have led to increased attention to the issues of packaging, producibility, and design standards. Previously, it has been generally assumed that smart pixel arrays should be distributed on regular square or rectangular grid arrays. For shuffle-based MINs, this Letter suggests that the smart pixels should not be arrayed on a rectangular grid and the smart pixel unit cell should not be a single pixel, or even a pair of pixels, but rather the kernel of a self-similar grid pattern.

The self-similar grid approach provides a much better match to MCM packaging technology than one based on a rectangular grid. The resulting design should therefore be significantly lower in volume and therefore cost.

References

1. A. Lohmann, W. Stork, G. Stucke, 1985 Opt. Comp. Mtg., WA3.

2. A. Lohmann, Applied Optics 25, No. 10, 1543 (1986).

3. S. H. Lin, T. F. Krile, and J. F. Walkup, Proc. SPIE, Vol. 752 (1987).

4. C. Stirk, R. Athale, and M. Haney, Applied Optics 27, No. 2, 202 (1988).

5. H. S. Stone, IEEE Trans. Computers, C-20, 153-161 (1971).

6. A. Sawchuk and I. Glaser, Proc. SPIE, Vol. 963 (1988).

7. M. W. Haney, Optics Letters, Vol. 17, No. 4, February 15, 1992.

8. C. E. Bauer, Proc. of International Sym. on Microelectronics, October, 1992.

9. B. Mandelbrot, The Fractal Geometry of Nature, Freeman, San Francisco, 1983.

3.2 Sliding Banyan Network

3.2.1 INTRODUCTION

The explosive growth in the Asynchronous Transfer Mode (ATM) equipment industry is just one indication of the ever increasing demand for high throughput, cost effective, broadband data switching networks. The high throughput demands of future systems will be driven by the growing number of nodes on any given network, the increasing bandwidth of data communications between nodes, and the desire to transmit video data over the same networks. Aggregate capacities in the Terabit/second regime will be required to meet the demand [1].

A common measure of interconnection difficulty in networks is the bisection width, defined as the minimum number of “wires” that must be removed to partition the network into two halves with identical numbers of processors [2]. Multistage interconnection networks (MINs) suffer from bisection widths that grow nearly linearly with the number of nodes. A banyan network is defined as a MIN with a unique path from any input to any output. Banyan-based MINs offer the potential for using simple self-routing algorithms that will be critical to the effective operation of high throughput switching circuits, where global control is impractical. Though theoretically powerful, banyan based switching architectures have been limited to fairly small network sizes owing to the high bisection widths of large networks. The link interconnections in a banyan network consist of perfect shuffles [3], or interconnections isomorphic to the perfect shuffle. These interconnection patterns do not lend themselves to implementation with traditional metallic interconnection techniques. The global interconnection topology of banyans leads, in VLSI approaches, to inter-node communications performance limits in speed, crosstalk, and power consumption. Very often the limitations of electronic banyans have caused designers to give up on exploiting the banyan network structure altogether, and adopt wholly different topologies, such as mesh networks, which have but simpler interconnection requirements, but much higher switching complexity. These approaches are not suitable for high throughput packet switching applications due to the large amount of buffering and contention control required. There is a need to use MIN approaches with new technologies that overcome the metallic interconnection bottleneck.

In this paper, the Sliding Banyan (SB) [4,5], a new MIN switching architecture that uses optoelectronic "smart pixels" and free-space optical interconnections to overcome the limitations of electronic interconnections, is described and evaluated. The SB provides a significant reduction in the number of switch, control, and interconnection resources that would be required in an equivalent all-electronic approach. The resource reduction stems from a novel partitioning of the resources – achieved by spatially interleaving the stages of the switching network in a way that is possible only with 3-D optical interconnections. The interleaving is possible because each stage in the shuffle-based SB multistage interconnection network requires the same shuffle link pattern. With interleaving, each stage's identical I/O pattern is slightly shifted from those of the other stages. A *single* optical system, suitably configured, can thus be used to interconnect all of the MIN's stages *simultaneously*. The SB uses a pipelined destination-tag self-routing approach within a deflection routing strategy. Since each node is physically co-located with all of its sister nodes at each stage, successfully routed packets may exit the network immediately, at whatever stage they finally arrive. This is the key feature of the SB. Deflected packets are effectively routed to a new banyan that has "slid" in time to accommodate those packets' needs. The result is that packets are removed from the network as rapidly as possible, leading to an overall low blocking probability and high resource utilization.

As a preface to the SB architecture description, Section II provides background on banyan networks and the optical shuffle-connected approach that are the key elements of the SB concept. The SB architecture is then detailed in Section III. In addition to the SB architecture description, Section III contains the results of simulations, analysis, and experiments that demonstrate the SB's important features. The simulations and analysis show that the blocking performance of the SB compares favorably with other approaches under fully loaded permutation traffic. In fact, the SB is shown to get within a factor of three of the theoretical minimum number of switching resources, despite using a simple self-routing control strategy. Furthermore, the results of experiments, in which vertical cavity surface emitting laser (VCSEL) and detector arrays were used to simulate future smart pixel I/O, indicate that an optical shuffle interconnection system based on conventional refractive elements is feasible. Section IV contains a discussion of the smart pixel technology needed to implement the SB network and the anticipated performance parameters of a future implementation. The Conclusion, contained in Section V, summarizes the key features of the SB network.

3.2.2 BACKGROUND

A. Shuffle-based Banyan Networks

Figure 1 depicts a shuffle based banyan network for N = 16 nodes. A banyan network connects any given input/output node pair with a unique path. There are typically $\log_k N$ stages, with each stage consisting of a permutation link pattern, such as a shuffle, and a set of k×k crossbar switches. In many banyans, the unique path between input and output is determined by following a simple self routing algorithm. This algorithm is called a destination-tag algorithm because it has the advantage of requiring only the destination address – the routing algorithm is independent of source address. Self routing eliminates the need for external control of the switching elements and offers the means to the high

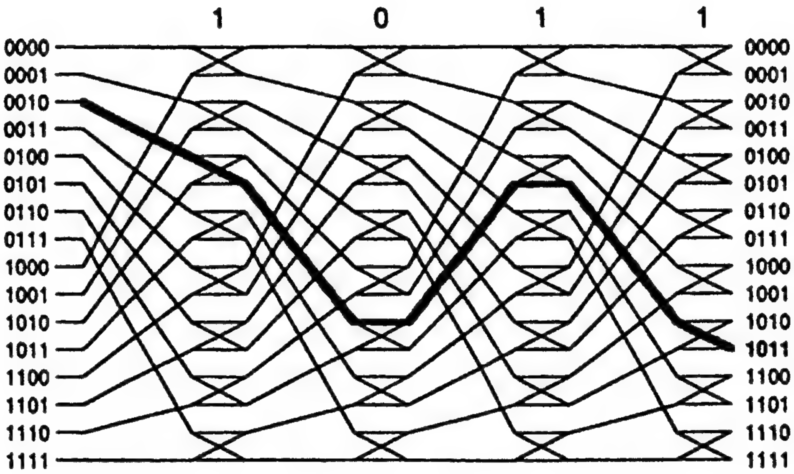


Figure 1. Shuffle based banyan depicting destination tag self-routing approach.

throughput desired in future switching circuits. Figure 1 illustrates this approach for a banyan comprised of perfect shuffle permutations and 2×2 switches. The self routing algorithm is performed on the output address (located in the packet header in ATM switching) as follows: beginning with the most significant bit (MSB) of the output address, inspect one bit at each stage – if the bit is a one, exit the stage on the lower node; if its a zero, exit on the upper node. Since the perfect shuffle performs a bit rotation on the address, this algorithm will work for any destination and is independent of the source address. Figure 1 shows this algorithm for a 16 node banyan.

In a single banyan architecture, a blocking error occurs when 2 (or more if k>2) packets wish to continue to the next stage using the same connection. The connection can only transmit one of the packets, so the other(s) must be lost, or blocked. A blocked packet has no method for recovery in this single banyan architecture. The blocking probability of a banyan increases as the load (number of active inputs) increases.

Redundant banyan architectures were proposed to overcome the internal blocking problem. Examples include replicated [6], dilated [6], and tandem [7] banyan networks, which all use auxiliary banyan networks to reduce the blocking probability to an acceptable level. Figure 2 depicts a tandem banyan (TB) architecture. In the TB, packets are not blocked, but rather tagged as having been routed incorrectly, and then passed through the remainder of the banyan. If two packets contend for the same output pin and one of them has been tagged as having been routed incorrectly, then the untagged packet

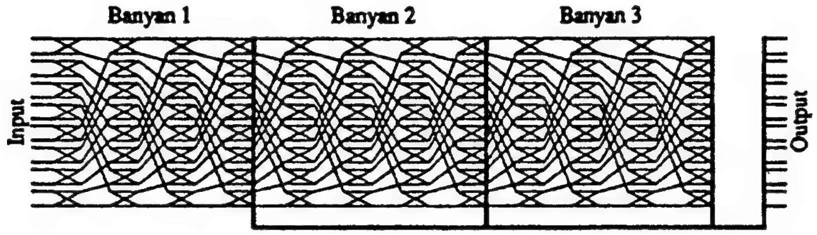


Figure 2. Tandem Banyan architecture.

would "win" the contention. If both packets have not been tagged, the winner could be determined by a random selection. At the end of one banyan, the packets which made it to their destination are removed from the network; the remaining packets then begin anew in a second banyan. Since some of the packets have been removed, the chances of blocking are diminished. Additional banyans are appended, in tandem, until an acceptable overall blocking performance is achieved.

B. Optical Shuffle Interconnections

To overcome the limitations of metallic interconnection, banyan based topologies can be implemented using free-space optics. An important step was the proposal of a free-space optical implementation of the perfect shuffle (PS) [8]. Several implementations of 1-D and 2-D optical shuffles were investigated [9-16]. All of these approaches use optics to effect the magnification and interleaving needed to perform the PS link pattern. Examples of 2-D shuffles are the Folded Perfect Shuffle [13], and the Separable Perfect Shuffle [9, 11].

A useful generalization of the PS pattern comes from partitioning the nodes into k equal sized groups and interleaving them. This is referred to as a k-shuffle; by this definition the PS is a 2-shuffle. The higher order shuffle does not avoid the high bisection width problem of the PS. However, the number of stages, and hence the switch latency, in k-shuffle based banyans is reduced to $\log_k N$, which is a strong function of k when $k<<N$. The price paid for fewer stages in a k-shuffle banyan (with $k>2$) is a need for more complicated k×k active self-routing switching elements.

One implementation of the separable k-shuffle uses two k×k lens arrays [16], a side view of which is depicted in Figure 3, for $k=4$ and $N = 16\times16$ nodes. In this arrangement each pair of lenses, from the two lenslet arrays, perform a unity magnification operation that achieves the desire k-shuffle pattern as shown. Figure 3 also depicts a grouping of the 16 smart pixel nodes into 4 subgroups in which the nodes are in close proximity to each other. This subgrouping concept places the nodes on a self-similar grid [17] rather than a regular square grid. As shown in the figure, the self-similar grid concept groups the smart pixels in a manner more amenable to packaging on separate OEICs and

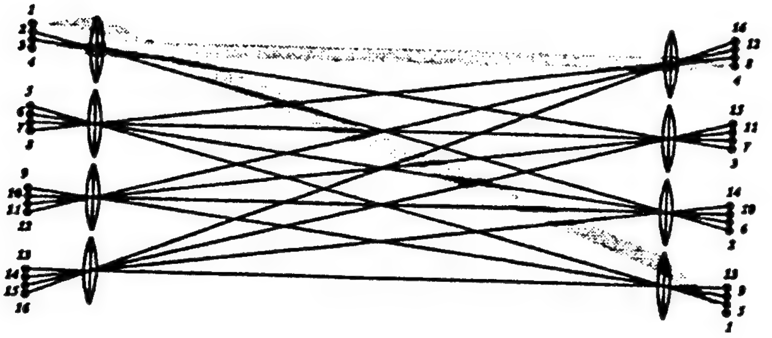


Figure 3. Side view of 4×4 shuffle system showing a 1-D 4-shuffle pattern arranged on self-similar grid layout.

increases the optical efficiency of the shuffling optics [17]. The symmetry of the optical k-shuffle depicted in Figure 3 is a key aspect that is exploited in the SB implementation described below.

3.2.3 SLIDING BANYAN ARCHITECTURE

A. Optically Interleaved Interconnection Topology

Interleaving of multiple shuffle stages was previously proposed to make better use of the shuffle optics' space bandwidth product (SBWP) and simplify the optical complexity by simultaneously using a single optical system for all stages in the MIN [18]. Figure 4 depicts the central notion of the interleaved topology used in the SB. Previously proposed MINs were comprised of physically separated stages – essentially emulating the traditional VLSI approach by replacing inter-chip and inter-board metallic interconnections with inter-chip free-space optical interconnections. Such a scheme, implemented with 2-D optical PSs [9,11,13,14], is depicted in the top half of the figure. This approach shows promise for overcoming the massive interconnection requirements between MIN stages, but has some implementation difficulties that stem from the physically separated multistage topology and the lack of compatibility with broad area multichip packaging conventions. The multistage implementation shown in the top half of the figure requires one array for each stage.

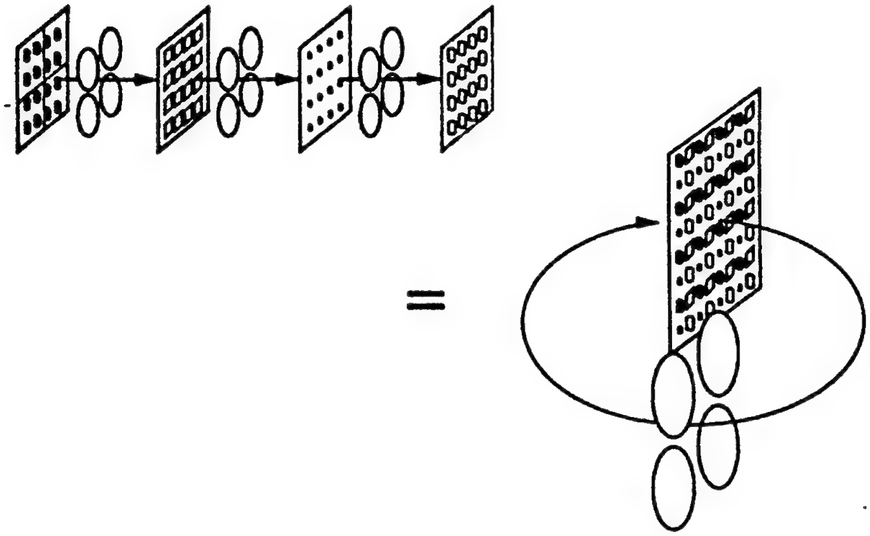


Figure 4. Optically interleaved shuffle based MIN topology.

Inspection of the network depicted in top half of Figure 4 reveals that the inter-stage interconnections are all identical and are shift invariant within the field of view (FOV) of the imaging optics. Therefore, with slight physical offsets of the I/O of each stage, multiple parallel interconnections can be implemented in an *interleaved* fashion, with a *single* optical system. This is schematically shown in the bottom of Figure 4, where a single optical system effects all of the required interconnections *simultaneously*. With this topology, the switching resources are distributed *laterally* across a single physical plane, rather than longitudinally across several planes. Since all of the stages have been collapsed onto a single plane, the bisection width implemented by the optics has been increased by the number of stages being implemented. To handle the increase in I/O resources, the plane on which the smart pixel array resides will be a PC board or multi-chip module (MCM) package that can accommodate an array of pixel optoelectronic integrated circuits (OEICs) that is large enough to contain all of the smart pixel resources. As discussed below, for a network with ~1024 nodes, such a system will likely be ~10-20 cm across. The associated optics, therefore, will consist of one or two *macro*-lenslet arrays, in which the size of individual elements correspond roughly to the size of an OEIC chips (e.g., 1-3 cm^2).

As discussed in the next section, the SB will require the equivalent of several banyans in stages to achieve the desired low blocking probability– e.g., for N=1024, approximately 30-50 stages will be needed. The number of stages that can be interleaved in this fashion is theoretically bounded by the SBWP/N. For example, consider a 1024 node system in which the PEs are arrayed in a 32×32 array. A typical high quality imaging system will have a SBWP $>10^6$, meaning that many more stages than required could theoretically be interleaved in this manner and maintain good isolation. A more practical limitation on the number of stages, however, is obtained from the real-estate constraints of the smart pixels and the related heat dissipation issues that determine the closest separation of emitter elements on the array. For example, typical air cooled IC's are limited to a few watts/cm^2 of power dissipation. If the optoelectronic elements of the smart pixel are assumed to dominate the power dissipation, and the power needed for a smart pixel link is of the order of a few milliwatts, then the number of emitters/detectors will be limited to several hundred/cm^2. This density limitation will determine the number of smart OEICs used, the number of nodes/OEIC, and the number of stages/node located at each smart pixel site.

In the schematic depiction of Figure 4, the shuffling optics interconnect the front of the plane to the back of the plane with the smart pixel logic connecting the two sides through the substrate. It is desirable to place emitters and detectors on the same side of the substrate, leaving the backside of this plane to function as a "backplane" and interface with electronic boards behind the optoelectronic backplane. A number of implementations of this approach can be considered; this paper focuses on an implementation based on the reversible k-shuffle interconnections, of the type depicted in Figure 3.

In the SB, each node in Figure 3 should be considered to actually be an identical cluster of optoelectronic elements, corresponding to all of the stages of the network. For example, in a 256 node SB switch, with 25 interleaved stages, there will be an array of 25 light sources interleaved with an array of 25 detectors located at each smart pixel location in the figure. Since the optical system is shift invariant within the field of view of each pair of elements, each of the rays in Figure 3 should be considered to be interconnecting the

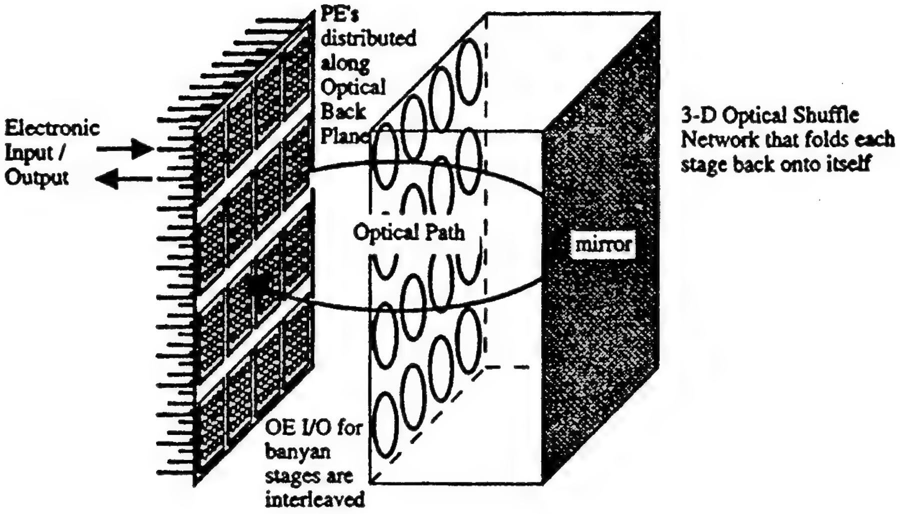

PE's distributed along Optical Back Plane
Electronic Input / Output
3-D Optical Shuffle Network that folds each stage back onto itself
Optical Path
mirror
OE I/O for banyan stages are interleaved

Figure 5. Single lenslet array k-shuffle interconnection approach. An array of processing elements (PEs) is interconnected through a shuffling retro-reflective optical system in an interleaved fashion.

cluster of sources at the input with the cluster of detector elements at the output, in a 1:1 manner. The second plane (e.g., the one on the right in Figure 3) could be used as an auxiliary active plane for routing purposes. However, a preferred and more compact variation of this concept is shown in Figure 5. Here a mirror is used to retro-reflect the shuffled image of the interleaved array back onto itself. A single lenslet array performs the k-shuffle interconnection in each dimension. When coupled with the self-similar grid array concept, the optical architecture depicted in Figure 5 uses a single lens for each OEIC. Thus each lens is simultaneously the input and output optical element for an OEIC.

The high resolution interleaved shuffle interconnection scheme depicted in Figure 5 demands lenses that provide wide field imaging with very low aberrations and good lens-to-lens (and hence OEIC-to-OEIC) registration. It is therefore likely that the required lenses will be multi-element and/or aspherical in design, with the possible inclusion of diffractive optical techniques. Fortunately, there is a wide body of applications that demand similar requirements, most notably high performance video camera and projection lenses. In the SB, the basic imaging system consists of a lenslet, a mirror, and another lenslet that is laterally offset from the first lens. The amount of lateral displacement is determined by the relative positions of the OEICs on the back plane. The SB, therefore, has several optical design issues to be resolved. These include: alignment, interleaved registration, distortion, focal length variation tolerances, VCSEL image resolution, and the retro-reflective folding of the optical system. These issues combine to determine the ability of the optical system to image the interleaved VCSEL array onto the interleaved detector array with good efficiency and low crosstalk. The alignment tolerances are determined by the size and spacing of the detectors and VCSELs. One compensating feature of the smart pixel design for the SB will be the use of monochromatic sources, such as VCSELs. Chromatic aberrations will therefore not be an issue.

B. Routing Control Approach

The physical collocation of stages afforded by the 3-D optics offers an important feature not practical with a VLSI implementation. Because the interleaving scheme allows all stages of a single node of the network to be in close proximity to each other, they may all reside on a single OEIC. This means that packets may be removed from the network *at any stage*, not just at the end of a fixed banyan, as in the TB. This is the essential feature of the SB architecture.

Figure 6 depicts the *unfolded* shuffle based SB architecture. The SB takes advantage of the constant availability of the output to reduce the network traffic quicker and thereby increase the networks performance. When a packet's route is blocked it is misrouted *once*, then it begins routing immediately. If the packet is not misrouted again, it will reach its destination in $\log_k N$ stages from the point of its deflection. This will be the end of this packet's banyan, which has slid to align with the misrouting incident, and the packet will be removed. With the SB, resources are not wasted by simply routing the misrouted packets to the end of the banyan; the rerouting begins immediately. After the first banyan, packets can physically leave the network at any stage.

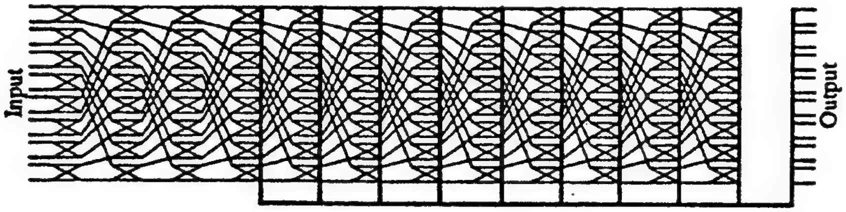


Figure 6. Shuffle based Sliding Banyan architecture (shown unfolded).

The Sliding Banyan routing strategy is possible only because the interleaved topology afforded by the 3-D shuffle optics provides co-location of the outputs for any stage of a given output node. Thus, only a single IC output driver is required for each node. If the unfolded network depicted in Figure 6 were to be implemented in VLSI technology, then an output driver for each node *and* for each stage would be required. In other words, each of the thick vertical lines in Figure 6, comprising N links for N nodes, would require a physical link and link driver of some sort. This would be totally impractical for a large network due to the large power consumption required for each of the output drivers. For example, consider a 1024 node SB with 40 stages, in a perfect shuffle (k=2) based banyan network. The first banyan requires $\log_2(1024)=10$ stages. The optically folded and interleaved topology of the SB requires just 1024 output drivers from the smart pixel plane. A VLSI implementation, in which the outputs from each stage reside on different ICs or boards, however, would require $(40-10)\times1024$ output drivers – a factor of 30 more than the optical SB.

This advantage is not without some added packet coding complexity. If packets are to be removed from the network at any stage, the packet must contain information about the number of stages correctly routed. When this number reaches $\log_k N$, the packet exits. Self-routing algorithms use a header with the destination address. The TB also requires a conflict bit to determine if the packet has been misrouted. Often the destination address will be rotated by each stage as it is routed; this way, the next stage need only inspect the first bit (if k = 2) to determine the switching. For the SB, the conflict bit would be replaced with a header containing the number of successfully routed stages, and the destination address would not be rotated. The number of successful stages is used to determine on which bit of the destination address the packet is to be routed. It is a simple inspection – if the number successful stages is *m*, then this bit of the destination is the determining one. Furthermore, the SB can give priority to those packets which had the highest number of successful stages in their history, i.e., are closest to their destination. This would prevent a packet, which had just begun rerouting, from interfering with a packet that is close to completing is routing through the network. The conflict bit is replaced by this priority number. Misrouted packets simply set this number to zero, then begin routing again.

C. Sliding Banyan Performance

The type of folded optical shuffle approach employed will determine the order (k) of the local cross-bar switches on the OEICs, The routing algorithm must then accommodate the local switching scheme within the smart pixel associated with the k adjacent nodes that must pass through the local switch. A digital simulation and analytical model have been developed for estimating the blocking performance of the SB and other similar networks, under various operating configurations and traffic conditions. Following are results which validate the SB in terms of blocking performance, latency, and switching resource requirements.

First, the TB was simulated on a 1024-node network with a 2-D (32×32) separable shuffle interconnection (k=4). The number of stages per banyan is n = $\log_4(1024) = 5$. The simulation tagged misrouted packets so that they would not interfere with any correctly routed packets. Randomly generated unity permutation traffic was used in the simulations. In this traffic pattern every input and every output is used exactly once; there is no output conflict. This is a standard type of traffic used in evaluating such networks. A typical plot of the number of packets remaining on the network versus the stage number is shown in Figure 7. Packets are removed only at the end of banyans; this is why the number changes only at integral numbers of five stages. Notice that the final banyan in vastly underutilized; it only routed one packet for this run. A total of 7 banyans were required to route the packets, so 35 stages were needed in all.

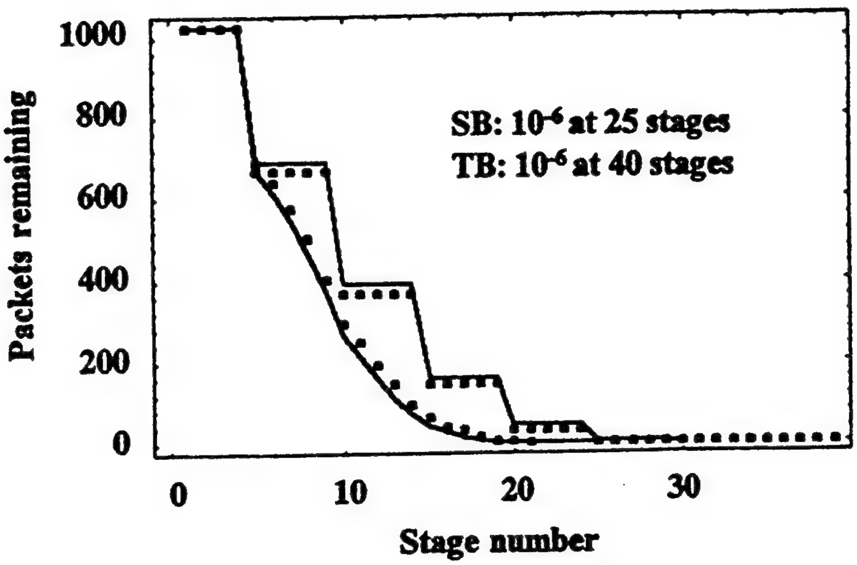


Figure 7. Performance of Sliding Banyan and Tandem Banyan networks. The solid lines show the simulated performance results, while the dotted lines show the analytical performance prediction results.

Next, the SB was simulated, using the identical unity permutation traffic pattern (originally generated in a random fashion) that was used to evaluate the TB. The packets were tagged with the number of consecutive successfully routed stages, and this number was used to prioritize the routing of any conflicts that arose. These results are also shown in Figure 7. Note that no packets are removed before the 5th stage, but then packets are removed at every stage thereafter. The packets removed in the 6th stage are those which were misrouted in stage 1, then had 5 successful stages. The network required 21 stages to route all 1024 packets, resulting in 14 fewer stages then the standard TB.

As a check on the simulated blocking performance results, a statistical model of the probability of blocking in the SB and TB was developed. This model is based on a modification to a banyan performance expression derived in [6]. The results are plotted in

Figure 7 along side of the simulation results, and show close agreement to them. Using the analytic approximation, the number of stages required for an arbitrary packet blocking probability (P_B), was determined; the results are plotted in Figure 8. These show that the SB maintains an advantage in number of stages over a wide range of operational P_B.

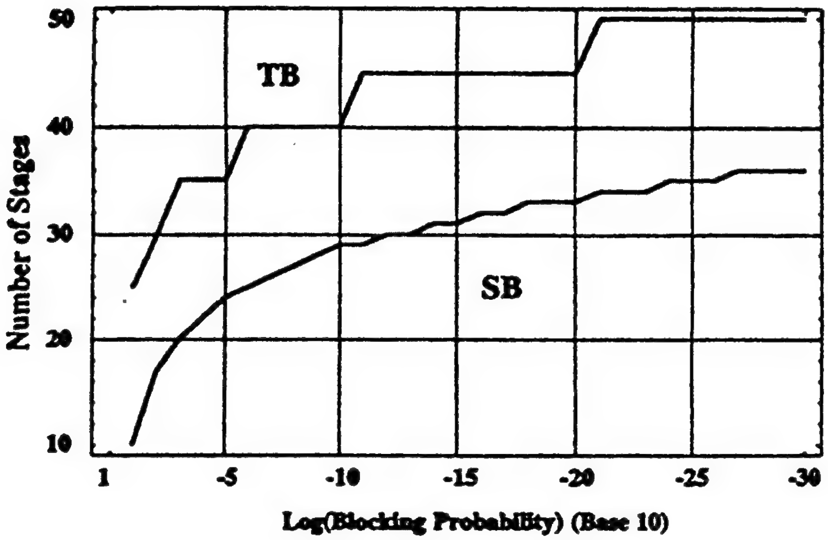


Figure 8. Number of network stages as a function of required packet blocking probability for the Sliding Banyan and Tandem Banyan Networks.

The analytical model used to generate the data in Figure 8 is based on the approximation that the probability that a packet survives the first m stages of a network composed of k×k switch elements, when the probability of a packet entering the first switch is p, is given by [6]:

$$p_m(k,m,p) = \frac{2k}{m(k-1) + \frac{2k}{p}} \qquad 1)$$

The probability that a packet entered this series of m stages and was blocked is the difference between p and p_m:

$$p_b(k,m,p) = p - \frac{2k}{m(k-1) + \frac{2k}{p}}$$

2)

Using this expression for packet survival, the analysis of a TB network is straightforward. Since unity permutation traffic is assumed, the initial probability (p) is 1, and one banyan's worth or $\log_k N$ stages is considered at a time. Using this iterative expression the probability of blocking in the ith banyan can be expressed in terms of the (i-1)th:

$$p_i = p_b(k, Log_k N, p_{i-1})$$

3)

In this manner the probability of blocking in one banyan is used as the input probability in the next. The load of the network is reduced until the probability of blocking of the final banyan is below the threshold required by an application, in this case 10^{-6}. *i* banyans or $i \times \log_k N$ stages are required to perform the routing with the requisite blocking probability.

The analysis of the SB architecture relies on Equation 1 as well, only it is a more complicated process. For the SB to be successful, packets which have completed the greatest number of consecutive successful routings must have priority over packets with fewer consecutive successful routings. When implementing the SB, a counter placed in the header indicated which bit of the destination address on which to decide – the higher the counter, the greater the priority. To simulate this mechanism all packets must be grouped in the network into $\log_k N$ groups. The designation of each group is the number of consecutive successful routings it has made ($P_\#$). All packets begin with zero successful routings. Again, unity permutation traffic is used, so the initial input probability is 1. The number of successful routings after 1 stage are determined as $p_m(k,1,1)$. This successful packet probability is placed in group 1 (i.e., one successful routing), and the probability of zero successes (group 0) is 1-P_1. At the next stage priority is given to those packets in group 1, P_2=p_m(k,1,P_1), but the packets with lower probabilities must contend with others like themselves, as well as the successful packets of group 1, so P_1=p_m(k,1,P_0+P_1). In this fashion every one of the probabilities of the $\log_k$N sized groups are calculated and the zero group is set to the remainder of all the packets (exclusive of those which successfully completed $\log_k$N stages and were removed). Thus, the manner of computing the probability of packet existence at every stage is:

$$P_i = p_m\left(k, 1, \sum_{j=i}^{Log_k N} P_j\right)$$

$$P_0 = 1 - \sum_{j=1}^{Log_k N} P_j \qquad 4)$$

When all of these probabilities have dropped below the threshold (i.e., 10^{-6}) then the network has achieved this probability of blocking and the number of stages required to achieve this is determined. These analytical expressions are a very close approximation for the simulations run, as shown in Figure 7. In order to achieve a probability of blocking of 10^{-6} for a 4 shuffle (k=4) the TB required 40 stages whereas the Sliding Banyan required only 25. This is a savings of approximately 30%.

As a measure of resource utilization efficiency, the optical SB self-routing network may be compared with the Benes network, which is known to be the smallest network for which all permutations are realizable, but for which no pipelineable self-routing algorithm exists [19]. The equivalent Benes network requires 2($\log_k$N)-1 stages, or 9 in our example traffic above. The 25 stages for the SB needed to achieve very low blocking probability is within a factor of three to the Benes, yet has the critical advantage of self-routing.

D. Optical Module Experiments

Current chip placement technology provides the ability to align ICs to a registration accuracy of approximately 10 μm across a multichip substrate. This registration accuracy will be suitable for the multichip SB implementation. If the smart pixel OEICs are assumed to have, on each chip, sub-micron optoelectronic registration (comparable to modern photolithographic IC technology capabilities), then the dominant source of optical misalignment and loss of efficiency will come from the lenslet array itself. Ultimately, custom wide angle imaging optics will be used for the SB. However, the requirements of the SB optical interconnection module are not unlike those of existing wide angle video and projection lenses. Therefore, preliminary experiments were conducted to evaluate the resolution and registration capability of commercially available lenses for use in the SB interconnection concept.

In the experiments, the smart pixel emitters were simulated with Honeywell-supplied arrays of VCSELs, including a 4×4 array of 10μm VCSELs on a grid with a center-to-center spacing of 630 μm and a 1×32 array of VCSELs with a center-to-center spacing of 140 μm. The VCSEL arrays were precisely placed at the various positions of smart pixel OEICs in the SB backplane, as depicted in Figure 5. The smart pixel output detector array was simulated by capturing the VCSEL array imagery on a high resolution CCD camera array, precisely positioned at other OEIC positions in the smart pixel backplane of the test set-up. Pairs of lenses under test were positioned to emulate the shuffle interconnection lens positions depicted in Figure 5. The results were analyzed assuming 40μm detectors spaced on the same grid as emitter arrays.

Figure 9 shows the results of a registration and resolution experiment in which 3 collinear VCSEL elements, spaced by 630 μm, were imaged onto a CCD array with a lens array system consisting of f/1.5, 25 mm focal length miniature video camera lenses. The overlaid white outline squares indicate the size and precisely registered locations of evenly spaced 40 um detectors that would be part of the smart pixel. As shown in the figure, the off-the-shelf video lenses perform fairly well in this off-axis imaging system. The resolution across the FOV of the system indicates that most of the light emitted by the VCSELs would be captured by an appropriately positioned 40 μm detector element. Some blurring occurred at the widest angle position (primarily due to vignetting of the narrow VCSEL beam by the barrel of the lens mount). The inherent distortion of the imaging system leads to misregistration of the VCSEL images with respect to the correct detector positions (indicated by the square box outlines). At the widest angles the array's images are beginning to misalign with the target detector array patterns. This distortion becomes especially apparent for total fields of view greater than about 20 degrees, occurring when the VCSELs were placed at the extreme points of the input field (approximately 4.4 mm from the axis). As shown in Figure 9, registration errors of approximately 25 μm occurred for the widest angle VCSEL images.

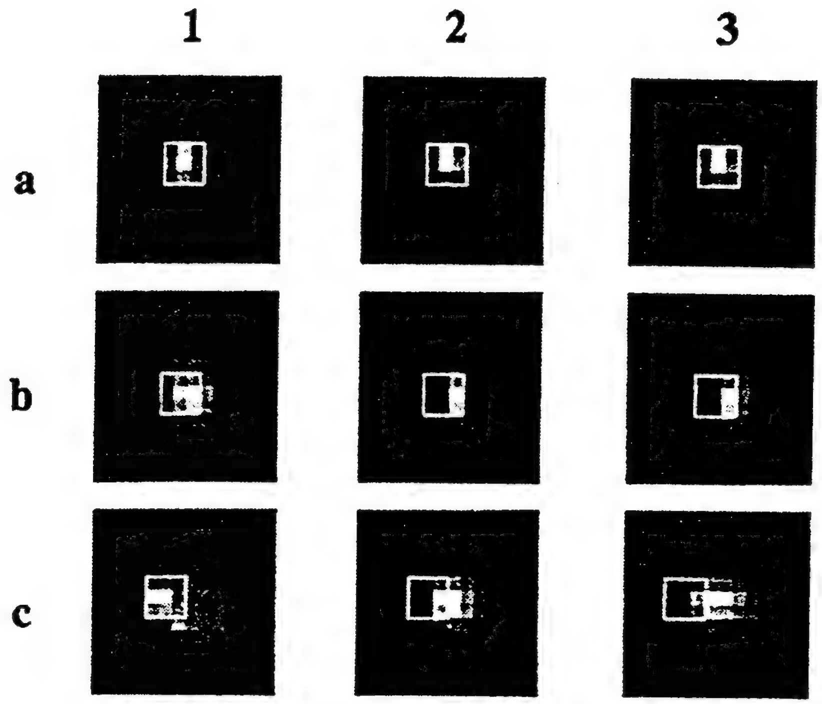


Figure 9. Focusing and registration data for 3 co-linear VCSELs separated by 630μm in the smart pixel plane. a: output for on-axis imaging, b: output for VCSELs centered ~2.5 mm from axis. (corresponding to ~5° off-axis), c: output for VCSELs centered ~4.4 mm from axis. (corresponding to ~10° off-axis). Box outlines correspond to the positions of properly registered 40μm wide detectors.

The focusing and registration results show good performance, despite the fact that the inexpensive test lenses were not selected to be precisely matched in focal length or other performance criteria. These results, therefore, suggest that better matching of commercially available lenses, or custom designed lenses, will provide the performance necessary for the SB optical module.

3.2.4 DISCUSSION

Several promising smart pixel technologies are now emerging as candidates for use in the SB packet switching architecture, including both emitter and modulator based approaches in monolithic and hybrid technologies [20]. At this stage of the study it appears that integrated emitter based (with either VCSELs or LEDs) smart pixels will more readily be incorporated into the envisioned optical system (as shown, for example, in Figure 5) than modulator based technologies.

In practice it is envisioned that packets will enter the SB switching fabric on a fiber optic or coax bundle, that interfaces to "line cards" stacked across the optoelectronic backplane. These boards then interface to the SB OEICs. For a 1024 node switch, consisting of 40 stages, there will be over 40,000 VCSEL/detector pairs distributed across the backplane. The power consumption is conservatively estimated to be 10 mW/smart pixel/stage, to include all electronic and opto-electronic power dissipation sources. Estimating power dissipation on a chip at a maximum of 2 W/cm^2 results in a maximum smart pixel I/O density of 200/cm^2. The SB architecture, consisting of 40,000 optical links, would then require ~150 cm^2 of OEIC chip area. A backplane of 20 cm×20 cm would have an OEIC fill factor of ~50%, which is consistent with practical MCM packaging.

3.2.5 CONCLUSIONS

Current all-electronic control and routing technology is not cost-effectively scaleable to the anticipated high throughput networks of the future owing to the fundamental limitations of metallic interconnections. The Sliding Banyan uses a fundamental advantage of free-space optical interconnections to reduce the switching and routing resources necessary in high throughput ATM switching applications. The novel free-space optical interconnection scheme provides the necessary high bisection width shuffle interconnection, while eliminating the need for large numbers of power hungry chip-to-chip drivers. Furthermore, the new 3-D interleaved topology, based on the rapidly maturing smart pixel technology, obviates the need for distributing the control and switching resources across numerous optical or electronic boards and instead provides a single backplane interface for the nodes of the switch. Preliminary experiments suggest that the high precision optical system needed to implement the Sliding Banyan can use existing high performance lens design techniques to achieve the need resolution and registration accuracy. Simulations and analysis show the Sliding Banyan approach to significantly reduce the resources required for a given blocking probability. The switching resources required to achieve blocking probabilities of 10^{-6} are within a factor of three of the Benes network, known to have the minimum number of stages for a non-blocking network. However, whereas the Benes network is totally impractical for high throughput ATM switching, owing to its lack of efficient routing control, the Sliding Banyan's fundamental advantage stems from its simple self-routing deflection control strategy, in which packets are removed from the network as soon as they find their way to their destination node. Self-routing control overhead is thus minimized in the Sliding Banyan. The combination of lowered switching resources and minimized control overhead of the Sliding Banyan topology provides an ATM switching architecture that is scaleable to aggregate bandwidths in the Terabit/second regime.

3.2.6 REFERENCES

[1] J. Hui, "Switching Integrated Broadband Services by Sort-Banyan Networks," *Proc. of IEEE*, Vol. 79, No. 2, pp.145-154, Feb., 1991.

[2] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures; Arrays, Trees, Hypercubes*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[3] H. S. Stone, "Parallel Processing with the Perfect Shuffle," *IEEE Transactions on Computing*, **C-20**, pp. 81-89, (1971).

[4] M. W. Haney and M. P. Christensen, "Optical Freespace Sliding Tandem Banyan Architecture for Self-routing Switching Networks," *Digest of the International Conference on Optical Computing*, pp. 249-250, August, 1994.

[5] M. W. Haney and M. P. Christensen, Free-Space Optical Sliding Banyan Network, *Digest of the OSA Topical Meeting: Photonics in Switching*, pp. 27-29, March, 1995.

[6] C. P. Kruskal and M. Snir, "The Performance of Multistage Interconnection Networks for Multiprocessors," *IEEE Transactions on Computers* **C-32**, No. 12, pp. 1091-1098, December, 1983.

[7] F. A. Tobagi, T. Kwok, and F. M. Chiussi, "Architecture, Performance, and Implementation of the Tandem Banyan Fast Packet Switch," *IEEE Journal on Selected Areas in Communications* **9**, No. 8, pp. 1173-1193, October, 1991.

[8] A. W. Lohmann, et al., in *Digest of the Conference on Optical Computing*, (Optical Society of America), Washington, D. C., paper WA3, 1985.

[9] A. W. Lohmann, "What Classical Optics Can Do for the Digital Optical Computer," *Applied Optics*, Vol. **25**, pp. 1543-1549, 1986.

[10] G. Eichmann, and Y. Li, "Compact Optical Generalized Perfect Shuffle," *Applied Optics*, Vol. 26, pp. 1167-1169, April 1987.

[11] S.-H. Lin, T. F. Krile and J. F. Walkup, "2-D Optical Multistage Interconnection Networks," *Proc. SPIE*, vol. 752, pp.209-216, 1987.

[12] K.-H. Brenner and A. Huang, "Optical Implementations of the Perfect Shuffle Interconnection," *Applied Optics*, vol. 27, pp. 135-137, Jan. 1988.

[13] C. W. Stirk, R. A. Athale, and M. W. Haney, "Folded Perfect Shuffle Optical Processor," *Applied Optics*, Vol. **27**, pp. 202-203, 1988.

[14] A. A. Sawchuk and I. Glaser, "Geometries for Optical Implementations of the Perfect Shuffle," *Proc. SPIE*, Vol. **963**, p. 270, 1988.

[15] M. W. Haney and J. J. Levy, "Optically Efficient Free-space Folded Perfect Shuffle Network," *Applied Optics*, Vol. **30**, No. 20, pp. 2833-2840, July, 1991.

[16] G. C. Marsden, P. J. Marchand, P. Harvey, and S. C. Esener, "Optical Transpose Interconnection System Architecture," *Optics Letters*, Vol. 18. No. 13, pp. 1083-1085, July 1, 1993.

[17] M. W. Haney, "Pipelined Optoelectronic Free-Space Permutation Network," *Optics Letters*, Vol. 17, No. 4, pp. 283-285, February, 1992.

[18] M. W. Haney, "Self-similar Grid Patterns in Free-space Shuffle/Exchange Networks," *Optics Letters*, Vol. 18. No. 23, pp. 2047-2049, December 1, 1993.

[19] F. Tobagi, "Fast packet Switch Architectures for Broadband Integrated Services Digital Networks," *Proceedings of IEEE* **78**, pp. 133-167, (1990).

[20] IEEE/LEOS Topical Meeting on Smart Pixels, IEEE Digest Catalog No. 94TH0606-4, July, 1994.

3.3 Sliding Banyan Network Performance Analysis

3.3.1 INTRODUCTION

There is an ever increasing demand for high throughput, cost effective, broadband data switching networks, as demonstrated by the explosive growth in the Asynchronous Transfer Mode (ATM) equipment industry. Future networks must handle thousands of high bandwidth channels, implying an aggregate capacity in the Terabit/second regime [1]. Electronic switching approaches, based on highspeed VLSI and associated packaging may not scale cost effectively above the 100 Gb/s regime [2]. Alternative technologies and architectures will be needed to meet the coming demand.

In this paper the performance of the Sliding Banyan (SB) network [3,4] is analyzed. The SB is a multistage interconnection network (MIN) architecture based on a new 3-D shuffle interconnection topology in which multiple stages' input/output (I/O) resources are interleaved (spatially multiplexed) on a common "backplane." With this topology, the critical I/O, switching, and control resources for a given node are placed in close proximity on the backplane, such that they are contained within the same optoelectronic integrated circuit (OEIC). With a suitable destination-tag self-routing control algorithm, the SB's physical arrangement permits rapid removal of correctly routed packets from the fabric. Initial simulations and analysis demonstrated the SB's advantages for permutation traffic [3,4] in which every input node is connected to exactly one output node, selected randomly. The SB was shown to provide a significant reduction in resources owing to its unique resource partitioning scheme. This reduction stems from the SB's topology and efficient self-routing control strategy, in which each packet is effectively routed through a banyan that has "slid" in the time domain to accommodate that packet's needs. This results in a significant reduction in internal contention in the switch, with a commensurate reduction in switching and interconnection resources.

The focus of this paper is on the performance of the SB under a realistic and demanding traffic load. Real systems characteristically have nonuniform traffic distributions. For example, in modern data communication systems, many nodes may require data from a single location (i.e., a server). In this scenario many packets may have identical destinations. This is, perhaps, the worst case scenario because the opportunity for internal contention is dramatically increased. Many topologies attempt to compensate for this type of internal contention through the use of redundant resources. The additional resources reduce the efficiency of the switch. However, as shown in this paper, the SB is particularly impervious to this inefficiency, owing to its ability to remove correctly routed packets from the switching fabric immediately. These benefits derive from the sliding time window made possible by the interleaved topology. This is a fundamental difference from other self-routing MINs in which, owing to limited I/O resources, packets are removed from the network only at the last stage, or possibly in a small subset of the stages.

Section 2 is a review of the SB topology, routing control algorithm, and optical interconnection scheme. The simulation used to validate the SB's performance are described in Section 3, along with the key results that include a comparison with other redundant banyan networks. Section 4 is a discussion of the results and their impact on resource requirements and scalability for the SB. In the Conclusion (Section 5), the implications of the enhanced performance on a physical implementation are discussed. Section 5 also includes a brief description of an experimental module, now under development to validate the optical interconnection subsystem of the SB concept.

3.3.2 SLIDING BANYAN NETWORK ARCHITECTURE

3.3.2.A Banyan Based MINs

Any network in which there exists a unique path from any input to any output is called a banyan network [5]. One particular type of MIN-based banyan consists of a set of $\text{Log}_k N$ stages, each containing N/k k×k crossbar switches interconnected in point-to-point butterfly or shuffle interconnection patterns. Figure 1 depicts a banyan network for N=8 nodes, with k=2. This type of banyan is particularly useful because simple destination-tag self-routing may be employed. As shown in the figure, each stage's k×k crossbar switch effectively sets $\text{Log}_2 k$ bits of the final destination address. When k=2, each stage sets one bit. The banyan depicted in Figure 1 is interconnected in a regular shuffle interconnection pattern – the pattern is identical between each stage of the banyan. Other interconnection patterns can achieve the same self-routing results. The type of interconnection describes the type of banyan. Some banyans have interconnections which differ between every stage (e.g., the butterfly), while others have the same interconnection pattern between all stages (e.g., the perfect shuffle (PS)[6]). In any case, the routing of the packet is independent of the packet's input location. This simple self-routing approach minimizes the control resources for packet routing, but suffers from possible internal packet contention – a fully loaded banyan cannot route all packets successfully. Packets need not be lost due to this, as there are always exactly k inputs and k outputs to each switch. A packet may simply be sent to an +incorrect address.

Redundant banyans have been suggested to overcome internal contention [7, 8, 9]. One such architecture, the Tandem Banyan (TB) [8], is depicted in Figure 2. In the TB unsuccessful packets are not blocked but rather "tagged" and misrouted to the end of the banyan. Packets which successfully arrive at their destinations are removed from the network, and the tagged packets' are reset, and sent into another banyan. The lower traffic on the next banyan makes it more likely that each packet will be successfully routed. Banyans are appended in this way until an arbitrarily low blocking probability is reached.

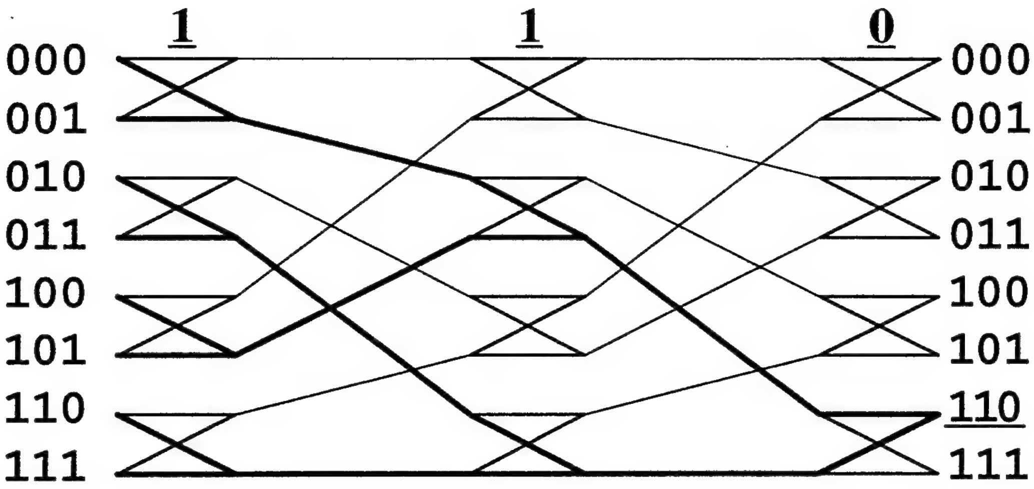


Fig. 1. Shuffle interconnected banyan network for N=8 nodes, consisting of Log_2N stages of N/2 2×2 switches connected by $(Log_2N)-1$ identical perfect shuffle interconnection patterns. The paths for destination tag routing for all inputs to destination 110 are highlighted. The switch settings, corresponding to the bits of the destination address, appear above each sequential stage. A packet is switched to the lower node if the bit is a 1, and the upper if it is a 0. This self-routing approach is independent of the packets input node, and its current location.

3.3.2.B Sliding Banyan Concept

The Sliding Banyan (SB) architecture takes advantage of a different type of redundancy to more efficiently handle internal contention. Instead of a set of discrete banyans, the SB is comprised of several "banyans worth" of identical perfect shuffle interconnections between stages, as depicted in Figure 3. A deflection routing scheme is used, instead of the TB's bit synchronous destination tag routing. In the TB, the first bit is examined at the first stage of the banyan, the second at the second, and so on. In deflection routing, used in the SB, the examination of bits is not tied to its location in a physical banyan. Instead, when a contention occurs, a "virtual" banyan slides to accommodate the immediate rerouting of the packet. Since all groups of Log_kN stages are interconnected with the same pattern, this virtual banyan remains unaffected by the location of the contention. This routing scheme allows for rapid removal of the packets from the network, but requires an output path at every stage of the network, instead of at the end of each banyan. This concept could become prohibitive for large networks if all the output paths were isolated output drivers. For example, if there are 1024 nodes (N=1024) with 60 stages, there would be 60,000 output drivers needed, for the implementation depicted in Figure 2. This is impractical for an all-electronic implementation owing to the prohibitively large number of required electronic output drivers and metallic interconnections (e.g., coaxial cable) [10].

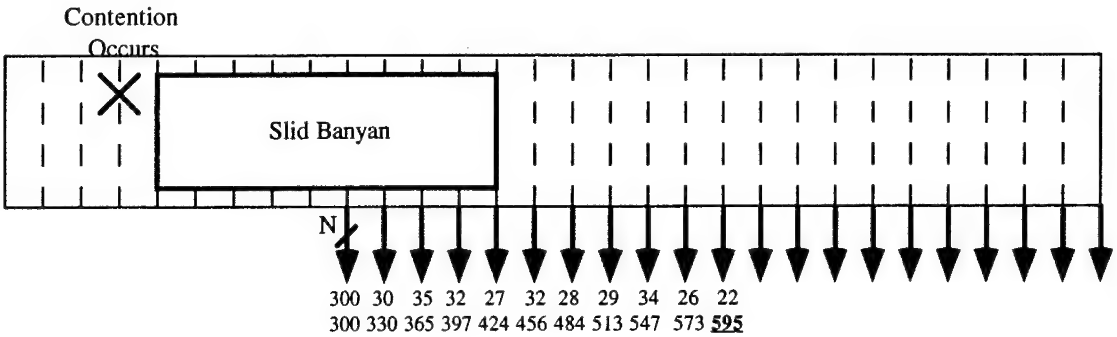


Fig. 2. The Sliding Banyan Concept: A collection of shuffle interconnected stages, with N outputs available at each stage after the first banyan's last stage. When an internal contention occurs, a virtual banyan slides to align with the next stage to accommodate the routing and removal of the lower priority packet in Log_2N stages from that point. The numbers indicate the current number and cumulative number of successfully routed packets at the end of each stage for the same simulated traffic that generated the numbers in Figure 2. The SB removes packets more rapidly than the TB.

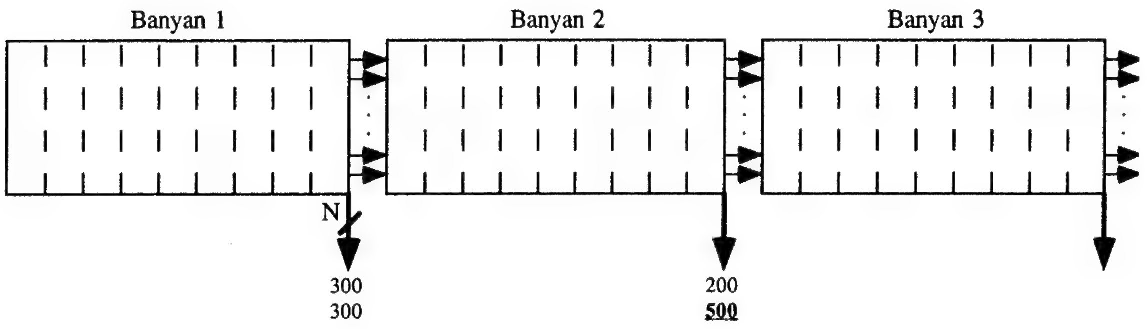


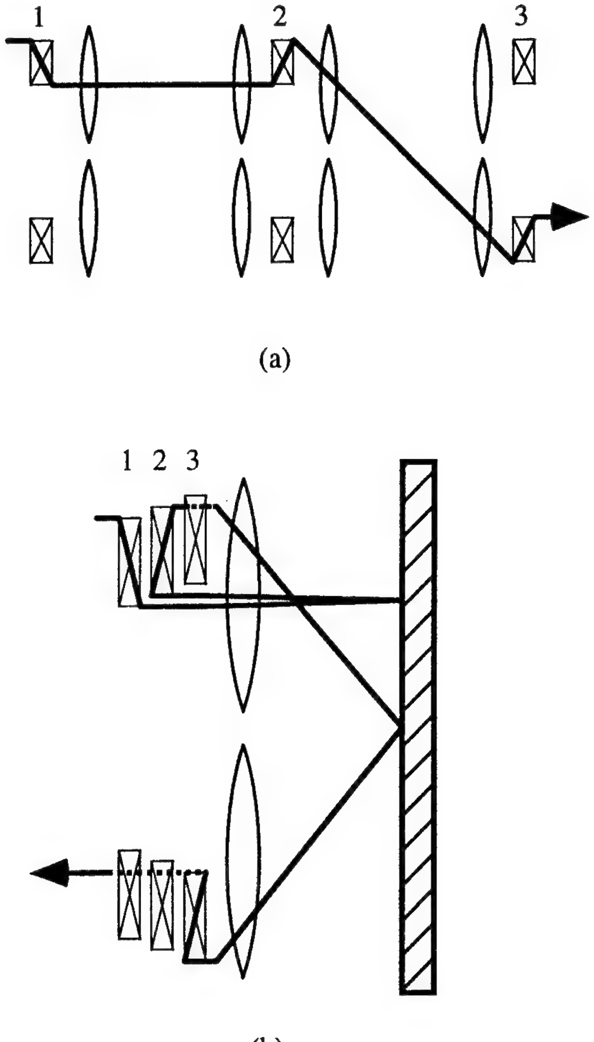
Fig. 3. The Tandem Banyan Concept: Banyans are 10 stages long, corresponding to N=1024. Successfully routed packets exit the switch at the end of each banyan. Unsuccessful packets are sent to the next banyan. The numbers indicate the current number and cumulative number of successfully routed packets at the end of each banyan for an actual simulation.

Since only a small fraction of the output drivers associated with any node can be utilized, it is advantageous to find a partition of the resources which removes this inherent redundancy in output drivers. This can be achieved by a partitioning scheme which spatially co-locates all stages of a given node, thereby allowing them all access to a much reduced number of output drivers. For a large switch, this is physically unrealizable in a VLSI-based all-electronic implementation due to the switching fabric's interconnection demands on the resource partitioning. However, as shown in the next section, using 2-D arrays of surface-normal optical interconnections in an interleaved shuffle removes the undesirable output driver redundancy.

3.3.2.C *Optical Sliding Banyan Topology*

The performance advantages of the SB stem from a new way of partitioning the resources – made possible by a retroreflective free-space optical shuffle interconnection approach. This 3-D shuffle interconnection approach provides a much higher bisection bandwidth capability than can be achieved by electronic packaging technologies. Several free-space optical shuffle interconnection approaches have been proffered [11-19]. The advantages of optical shuffle interconnects stem from the use of image interleaving techniques that provide point-to-point links across a 2-D plane of nodes, thereby avoiding the issues of physical wires and electrical signal coupling. Furthermore, the emerging technologies of low power vertical cavity surface emitter laser (VCSEL) and detector arrays permit the distribution of the I/O across the surface of the OEICs instead of around the perimeter.

Figure 4a depicts a schematic side view of an optically interconnected MIN for a square array of N=16 nodes. The shuffle links are implemented with physically separated optical systems – one for each stage. This is the typical approach considered in optically interconnected MINs. For simplicity, Figure 4a depicts a packet's route in one dimension. The optics shown implement a shuffle via a symmetric arrangement of lenses, requiring 2k lenses for each shuffle interconnect [19]. It has been pointed out, however, that the optical shuffle has a local shift invariance property that makes it possible to interleave



(a)

(b)

Fig. 4. (a) is a schematic side view of an optically interconnected MIN for a square array of N=16 nodes: The integrated plane of electronic switching and OE I/O are schematically depicted as 2×2 crossbars. An example path from the upper input port of switch I-1 to the upper output port of switch II-3 is shown. (b) is a schematic side view of the interleaved retro-reflective partitioning of (a). The stages of nodes, represented here as 2×2 crossbars, are interleaved. The same example path is depicted as in (a). The switching functions are expanded for illustrative purposes. In practice, all switches would be integrated into the single opto-electronic backplane. For large networks, this concept extends to one lens for each chip in the opto-electronic backplane, where each OEIC contains k×k nodes. Furthermore, each node contain one switch for each stage present in the Sliding Banyan.

multiple stages' I/O across a common plane and retro-reflect the shuffle links to all stages simultaneously [20]. Such an approach provides the ability to effectively pipeline multiple stages within a single optical system. Furthermore, the symmetry of systems like the one depicted in Figure 4a permit the shuffle optics to be implemented with a single array of lenses and a mirror [3]. This concept is illustrated in Figure 4b. The switching , control, and I/O resources of the multiple stages are repartitioned in an interleaved manner onto a single plane. All stages of a single node are co-located within that plane. If this co-location is provided on a single chip, then only one off-chip output driver is required for each node – a dramatic reduction from the requirements of Figure 4a. The new resource partitioning scheme utilizes a single macro-optical shuffle interconnection module, which is simultaneously used by all stages. The resources are therefore distributed laterally across a single backplane, rather than longitudinally across many planes. Such a planar distribution is amenable to implementation with conventional MCM packaging technology.

For the large (e.g., N = 1024 nodes) switching applications envisioned for the SB concept, the electronic switching and control functions are integrated with the optoelectronic I/O functions in an array of OEICs distributed across an optoelectronic backplane containing many such chips. Each OEIC will handle a subset of the nodes. For example, a possible configuration might be comprised of 16 nodes per OEIC, with an 8 × 8 OEIC array on the backplane – to give a total of 1024 nodes. If there are 50 pipelined stages in the SB network, then each OEIC requires 50 × 16 = 800 optoelectronic I/O. Several promising high-speed OEIC technologies, based on monolithic and hybrid integration of VCSELs and detectors with logic are now under development [21]. The SB's interleaved multi-chip optical shuffle approach combines the global interconnection advantages of optics with local low power on-chip electronic interconnections to effect the appropriate links and avoid redundant output drivers.

3.3.2.D Sliding Banyan Routing Control

High aggregate bandwidth (e.g., 0.1-1 Tb/s) switches will require distributed local routing control for scalability. Self-routing shuffle based networks, utilizing a deflection algorithm to route packets to their destinations, have been proposed [22]. The SB employs a variation of deflection routing that is based on simple destination tag routing. In this scheme the headers of each set of k (or fewer) packets entering a local k×k switch are examined to determine the proper switch settings. The simplest kind of contention resolution will occur for k =2. To simplify the explanation, k is assumed to be 2 in the following discussion, although the basic idea can be extended to higher order k-shuffle/exchange networks.

The basic banyan for the SB consists of Log_2N stages of perfect shuffle interconnections, each of which is followed by a set of N/2 2×2 cross-bar switches. Figure 1 shows how a packet would be routed in one of the SB's banyans. The packet is first shuffled from its original address to the first set of switches. If there is no contention then a switch controller examines the destination address's MSB and routes the packet to the "bottom" output of the 2×2 switch if its a 1 and the "top" output if its a 0. After the next shuffle stage, the next most significant bit is examined, and the switch set appropriately as in the first stage. At the final stage, the LSB of the destination tag provides the last switch setting for that packet. If 2 packets require the same output node of the switch, only the one with the highest priority is switched to that node, while the other is routed to the other node of the 2 × 2 switch and "flagged." A flagged packet must restart its destination-tag routing to begin anew at the destination address's MSB. The effect of this is that of a new virtual banyan being slid to begin at that packet's current location in the series of stages, as depicted in Figure 3. If no further contention occurs, the rerouted packet will complete its routing through the slid banyan and then exit the network at its destination node.

To determine which packets take precedence upon internal contention, packets are assigned a priority tag according to their current degree of success at routing. Packets which have nearly completed their routing have the highest priority. The degree of success is determined by how many consecutive stages the packet has successfully (i.e., without losing a switch contention event) traversed. If a conflict occurs between two packets, the header decoder routes the packet with the higher priority correctly and flags the other packet to begin rerouting at the next stage.

When a packet reaches its destination, a simple on-chip switch is required to route it out of the switching fabric, since all stages of the pipelined network are physically co-located. Packets are thereby immediately routed to there destination, without the burden of excess underutilized switching resources, complicated control, or a large number of power hungry highspeed output drivers. The distributed nature of the SB's routing control algorithm provides the means to high aggregate throughput, while the simple logic functions needed for the SB indicates that the overall overhead for control functions in the SB will be low.

The number of stages needed in the SB is determined by the overall blocking performance desired. Simulations and a blocking rate model were developed to evaluate the SB's performance under full permutation traffic conditions [4]. The model's predictions for blocking error rate followed very closely to the actual data achieved by compiling statistics over many simulation runs. Both perfect shuffle (k=2) and 4-shuffle configurations were evaluated for the TB and SB. For N=1024 and k=4, it was found that the SB required 30 stages to achieve an overall blocking probability of 10^{-12}, whereas the TB required 45 stages – a significant improvement in switching and interconnection resources and latency.

The key advantage demonstrated by the SB is the immediate re-routing of deflected packets. In the TB, when contention occurs, the deflected packet must be pushed along to the end of a banyan before it can begin its re-routing process. This is necessary due to the limited number of exit points the packet has available to it in the TB architecture. In the TB a packet can complete its routing only at the end of a banyan, therefore it can only begin routing at the beginning of a banyan. This results in many switching stages being utilized for passing a deflected packet along to the end of the banyan, instead of routing the packet to its destination. The numbers next to the exit lines of Figure 3 show how immediate rerouting of packets affects the load on the switching fabric for a typical simulation run. In the TB, no packets exit the network within banyans, whereas in the SB packets continually exit the network after the first banyan.

In the TB groups of $\text{Log}_k N$ stages (banyans) are appended to the end of the switching fabric until the desired arbitrarily low blocking probability is reached. This large commodity of stages can be wasteful, as the last banyan in a network seldom has more than one packet on it. The SB adds single stages as necessary for acquiring an arbitrarily low blocking rate. In this way, the SB maintains a resource advantage over the TB for all blocking probabilities.

3.3.3 AREA-OF-INTEREST TRAFFIC ANALYSIS

The analysis of permutation traffic provided a good first indication of the advantages of the SB. However, more realistic nonuniform traffic patterns, such as "area-of-interest" (AOI) traffic, are more challenging to switching performance and must be evaluated. AOI traffic patterns result when many packets are routed to the same destination. This destination may be several contiguous nodes, or simply a single node. For AOI traffic, it is assumed that output buffering is used such that all packets that are not lost within the switch will be successfully passed on to the output – so output contention is avoided. When a packet reaches its destination switch it is allowed to exit, even if the final switch has contention. However, since many packets may have the same output destination tag in their header, internal contention may diminish the overall performance of the switch.

For any given fixed switching architecture, AOI traffic will likely increase the overall probability of blocking for the switch. The selection of an architecture should be driven by the efficiency and robustness to the variations in traffic that may be encountered. A robust switching architecture will have very little change in efficiency or resource utilization under variations in traffic patterns.

To determine robustness and efficiency the average number of stages necessary to successfully route all packets was calculated for various AOI traffic patterns. The performance of the SB was compared with that of various types of TBs. From these data, the total amount of switching and interconnection resources was estimated and the overall resource utilization deduced. The program simulated the routing algorithms of the various switches, measured how many stages are needed to successfully route all of the inputs, and then compiled performance statistics for many thousands of runs. In particular, the SB was compared to several replicated versions of the TB under AOI traffic between 0 and 3 %. In this case a TB with a replication factor of q was simulated by q TBs with 1/q traffic load for each.

The simulated AOI traffic was generated as follows. First, a random permutation of the input set was generated. Then a fixed percentage (between 0 and 3%) of those addresses was set equal to a single randomly picked output node. The "area" consists of a single node in this case. The same traffic pattern was presented to the SB and various replications of the TB. Figure 5 represent the results of 1000 simulated runs of the SB and TB under AOI traffic. The mean number of stages required is plotted, with error bars indicating the standard deviation of the data. The TB data have an apparently larger error due to the fact that the possible number of stages were integral groups of Log_2N stages.

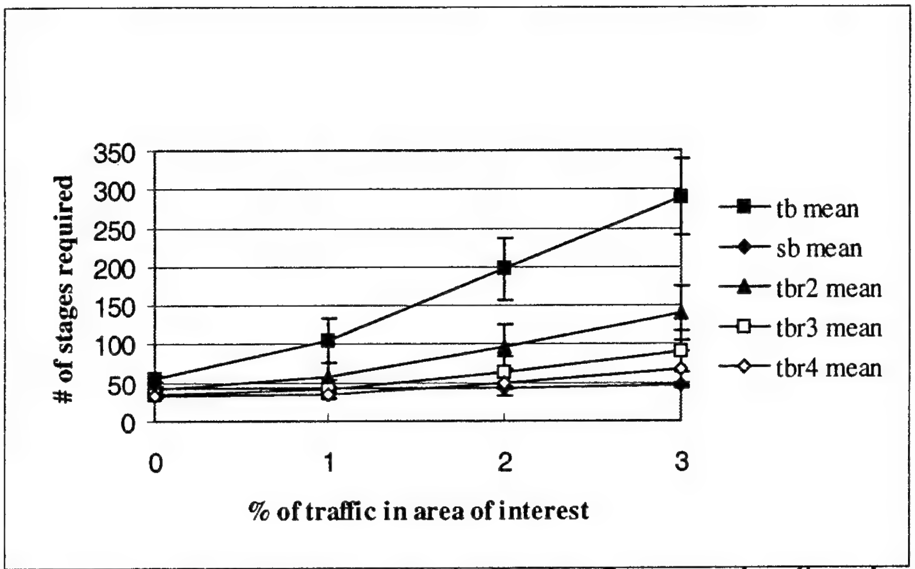


Fig. 5. Comparison of Sliding Banyan with the Tandem Banyan and replicated versions of the Tandem Banyan for area-of-interest traffic. The replicated Tandem Banyans have 2, 3 and 4 Tandem Banyans in parallel with 1/2, 1/3, and 1/4 of the traffic load, respectively, in each simulated run. The vertical axis represents the number of stages required to route all 1024 packets.

As shown in Figure 5, the AOI traffic has very little effect (less than 3-5 additional stages, or about 10%) on the SB. However, AOI traffic significantly increases the number of stages of a single TB – requiring well over 200 more stages for 3 percent AOI. The replicated TBs suffers less from this effect. The replicated TBs achieve this by decreasing the length (latency) of the switching architecture at the expense of its breadth. While this replication has a beneficial effect on the number of stages required for this traffic pattern in the TB, each stage now has several times the number of switching resources as a single stage of the TB or SB. These added resources are accounted for by normalizing the data in Figure 5 by the total number of switching resources. The replicated TBs have, in effect, traded resources for decreased latency. Yet, in ATM switching, relative latency becomes an issue only when it approaches the time length of a packet; if the latency is less than a packet length misordering of the data is avoided.

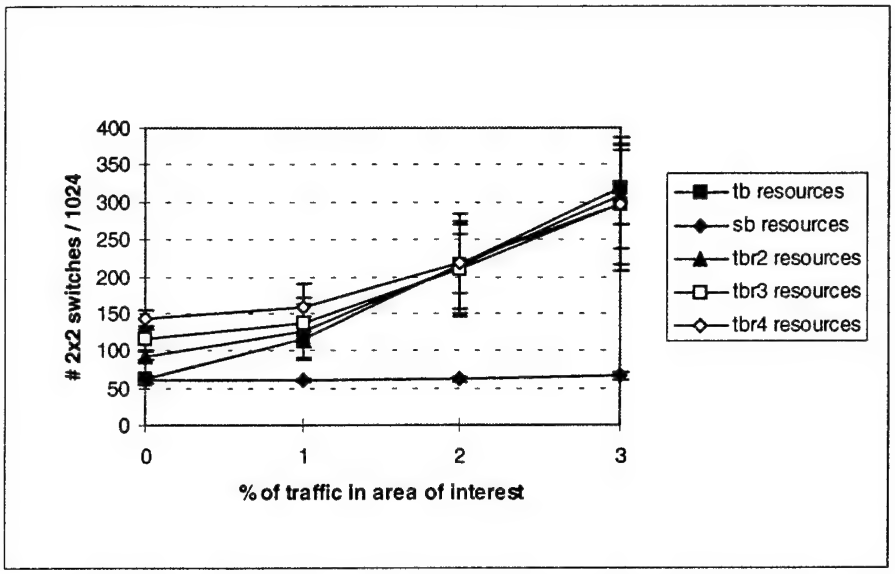


Fig. 6. The data of Figure 5 normalized to take into consideration the redundant resources of replicated stages and output switching. The vertical axis represents the switching resources required to route all 1024 packets.

Figure 6 shows the results of such a normalization of the data. These data show that while replications of TBs improve latency, the overall switching resources remain similar. The switching resource requirements grow at an undesirable rate under increasing AOI traffic, independent of the degree of replication. In this analysis, switching requirements also include output switching resources, i.e., those necessary for removal of packets. In the SB this occurs at every stage of the network, whereas in the TBs it only occurs at the end of each banyan. While the SB pays a heavier penalty in switching to exit the network, it is insignificant compared to the penalty paid by the redundant resources of the TB. In fact, the redundant versions of the TB perform equally poorly under this traffic pattern when resource utilization is examined. The resource requirements for the SB remain nearly constant over all traffic conditions studied.

3.3.4 4. DISCUSSION

The simulation results clearly show that the SB is relatively immune to AOI traffic effects while the TBs have great difficulty with AOI traffic, with any replication scheme. This difficulty is due to the TBs' bitwise synchronous routing scheme, the first bit is examined at the first stage, the second bit at the second and so on. If many packets are destined for the same area, most of their bits will be identical – if they are destined for the same node, as in the above simulations, the addresses are identical. After deflection occurs, the remaining packets all collide again and again on the same stages of each successive banyan, letting only a few pass each time. Recalling that 1% of 1024 node network is approximately 10 packets, and only two at most can succeed at each banyan, it is not surprising that this traffic pattern is difficult for the TB.

Replicated TBs reduce the latency of packet routing, allowing several times more AOI packets to be routed in each banyan. This somewhat reduces the problem of bitwise synchronous blocking in AOI traffic. This recurring bitwise synchronous blocking is similar to a problem faced in the ALOHA communications protocol [23], in that two packets may continually interfere with each other, and prevent successful routing. The accepted solution to this problem for ALOHA is to introduce a random delay when a potential blocking occurs. This de-synchronizes the packets and allows them to continue without continually colliding. This is analogous to how the SB avoids bitwise synchronous blocking. When a packet is deflected in the SB it begins re-routing immediately. This aligns the bitwise decoding of the packets header with its current stage (where the error occurred). This action spreads the packets headed for the AOI among several non synchronous (i.e., non interfering) banyans.

Similar simulation comparisons resulted between the SB and TB for different numbers of nodes. For example, it was found that whether N = 512, 1024, or 2048, the same relative efficiencies were obtained. This therefore shows that the SB's resource efficiency performance scales well with network size relative to the various TB variations tested.

Since the simulations provided a count of the actual switching and interconnection resources necessary to achieve good blocking rates, the absolute resource efficiency of the

SB can also be characterized. For example, it is well known that the Benes network uses (Log_2N)-1 stages, which is the minimum number of MIN switching and interconnection resources needed to achieve a rearrangeably non-blocking network [5]. However, the Benes has been found to be entirely impractical for large (many nodes), high throughput applications due to limitations on the necessary control algorithm. The implementation of a non-blocking network in the Benes concept requires a global control algorithm. No efficient (non-recursive or linear) algorithm is known for setting the switches. Furthermore the Benes has no provisions for handling AOI traffic. The SB simulation results above show that typically less than $5Log_2N$ stages are required to successfully route all packets with high AOI traffic. With the additional overhead of output switching, the total switching resources remain within a factor of 4 of the Benes network. Yet it maintains the critical practical advantages of distributed self-routing control and the ability to handle nonuniform traffic.

To investigate the algorithmic efficiency of a MIN architecture, the switching resource efficiency may be examined. In the SB and TB switching resource usage falls into three categories. A switching resource was either used to correctly route a packet, used to pass a packet to the next stage to later be routed, or was not used at all. Figure 7 plots the average percentage (over 1000 random simulation runs) of switching resources utilized for correctly routing packets in each of the networks. The SB remains efficient over the range of AOI traffic patterns, whereas the TB and its replications become very inefficient. The ideal would be 100% switching utilization efficiency. The Benes network, the minimum size re-arrangeable nonblocking network, would be 100% efficient for permutation traffic under this definition, but, as mentioned above, is impractical due to control issues and the fact that it cannot effectively handle AOI traffic.

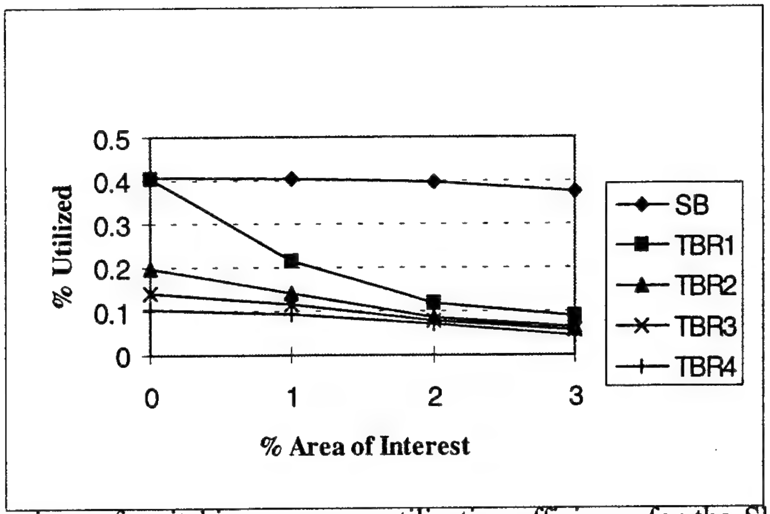


Fig. 7. Comparison of switching resource utilization efficiency for the Sliding Banyan, Tandem Banyan, and several replicated versions of the Tandem Banyan. Only switches which contribute to the correct routing of a packet counted for this measure. The Sliding Banyan maintains high switch resource utilization efficiency for a wide spectrum of nonuniform traffic.

The analysis in the previous section focused on the switching and interconnection resource requirements of the SB concept. However, a complete analysis must address all types of resources necessary to implement the SB. The resources necessary to implement any switching concept are broken down into three general categories: switching, interconnection, and control. The switching resources are further broken down into internal switching (routing of packets) and output switching (removal of packets). The interconnection resources are also further sub-divided into drivers and links. Here, drivers refer to electronic or opto-electronic inter-chip drivers. Inter-chip drivers are assumed to be a dominant source of resource requirements because intra-chip interconnections are typically much lower in power consumption. The link resources are comprised of physical inter-chip and/or inter-board metallic paths, such as high speed lines and coaxial cable, needed to implement the desired shuffle/banyan link patterns. For free-space interconnection systems, such as the one employed in the SB, the link resources are comprised of the optical elements and free-space volume necessary to achieve the required shuffle links.

For good scalability a switch concept must exhibit reasonable resource growth requirements with throughput for all of the main types of resources delineated above. Table 1 summarizes the growth requirements for the SB. In the SB, the internal switching resources are the actual local electronic switching elements necessary to affect the SB routing algorithm. The switching requirements for the internal switching is $5(2\times2)(N/2)Log_2N$. This corresponds to 5 banyan's worth ($5Log_2N$) of stages, each with N/2 switches of complexity 2×2. The coefficient 5 comes directly from the blocking probability; a different blocking probability would simply lead to a different coefficient. Previous studies have shown that a small increase to this coefficient results in a dramatic reduction of the probability of blocking. For example, for a change from 10^{-6} to 10^{-30} blocking probability, the coefficient increases from 5 to 7. The minimum number of switching resources is $(2\times2)(N/2)((2Log_2N) - 1)$, for the Benes network – but this requires complex control.

Table 1. Resource requirements for SB with $\sim10^{-6}$ probability of blocking.

RESOURCE	REQUIREMENT
Switching: Internal	$5(2\times2)(N/2)Log_2N$
Switching: Output	$4NLog_2N$
Interconnects: chip output drivers	N
Interconnects: free-space links	$N\ (5Log_2N-1)$
Control	$\propto N$

The output switching resources in the SB stem directly from the ability to remove packets at any stage after the first banyan. Therefore, the output switching requirement is $4NLog_2N$. This corresponds to adding N output paths to 4 banyan's worth of stages to allow packets to be switched out of the switching fabric. The number 4 is directly tied to the length of the network, it will always be 1 less than the coefficient for internal switching, since the first banyan requires no internal output paths.

The unique partitioning of the SB reduces the number of output drivers to N, one for each node of the network, whereas an approach with physically separated stages (as would be required in a VLSI implementation), would require $5NLog_2N$. The optical shuffle link requirements is equal to the product of the number of SB stages (minus 1) and N nodes since each pair of stages has N interconnects between them.

Practical limitations on the interleaved shuffle interconnection approach are determined by the quality of the optical elements and the physical optoelectronic I/O spacing requirements on and between the smart pixel OEICs [24]. Preliminary experiments with VCSEL arrays and wide angle imaging shuffle lenses have validated the basic retroreflective approach [4].

Finally, the distributed control of the SB allows a nearly linear growth in control resources as only one processor per node is required to implement routing for all stages for that node. The use of a single control processor for each node is achieved via the unique 3-D topology of the SB architecture – without the co-location of stages, this linear growth would not be feasible.

3.3.5 CONCLUSION

The realization of a SB switch with .1-1 Tb/s aggregate throughput [25] will hinge upon research in three related areas. The first area concerns the selection of the smart pixel technology, distributed across many OEICs on the SB's backplane, in which the required active resources will reside. Several promising monolithic and hybrid technologies in integrating electronic logic with arrays of emitters and detectors are rapidly emerging [21]. All of these approaches are pushing toward the VLSI density of logic required for the SB's smart pixels. Previous estimates of the required optoelectronic I/O density on the smart pixel OEICs [5] of several hundred per cm2 appear to be within reach of the smart pixel developers.

The second area concerns the design and optimization of the control algorithm and its implementation with a smart pixel technology. Several variations on the basic SB control strategy are currently being evaluated. These variations include an analysis of the degree (k) of the local switch. The analysis thus far has usually assumed k=2, providing the simplest SB node functional logic. However, higher order shuffle approaches will reduce the number of stages by approximately a factor of k-2, thereby reducing significantly the number of stages and optoelectronic I/O at the expense of more complicated nodes. These trade-offs must be evaluated to determine the optimum SB smart pixel configuration.

The last area concerns the experimental validation of the retroreflective interleaved optical shuffle interconnection approach. An experimental SB optical module, similar to the one depicted schematically in Figure 1, is currently under development [26,27]. It consists of a 4×4 array of identical high quality miniature projection lenses that have a wide and flat field-of-view, with high resolution. Preliminary testing of the experimental module shows that it will be capable of implementing a full banyan interconnection pattern for a 256 node array.

The AOI traffic analyses described in this paper demonstrated the benefits of the SB topology. Significant improvement over other types of redundant banyan approaches was demonstrated in total switching and interconnection resources required, as well as in latency. The performance enhancements of the SB stem from its unique partition of resources in a new 3-D optically interconnected topology in which a simple self-routing algorithm effects the rapid routing and removal of packets from the fabric. The SB's robustness to nonuniform traffic and its linear or nearly linear resource growth with the number of nodes suggest that this concept will scale well with network size.

3.3.6 REFERENCES

[1] J. Hui, "Switching Integrated Broadband Services by Sort-Banyan Networks," *Proc. of IEEE*, Vol. 79, No. 2, pp.145-154, 1991.

[2] T. Egawa, K. Yukimatsu, and K. Yamasaki, "Recent Research Trends and Issuesr in Photonic Switching Technologies," *NTT Review*, Vol. 5, No. 1, pp. 30-37, 1993.

[3] M. W. Haney, and M. P. Christensen, "Optical Freespace Sliding Tandem Banyan Architecture for Self-routing Switching Networks," *Digest of the International Conference on Optical Computing*, August, 1994.

[4] M. W. Haney, and M. P. Christensen, "Sliding Banyan Network," Submitted to: *Journal of Lightwave Technology*, August, 1995.

[5] F. Thomson Leighton, Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[6] H. S. Stone, "Parallel Processing with the Perfect Shuffle," *IEEE Transactions on Computing*, **C-20**, pp. 81-89, 1971.

[7] C. P. Kruskal and M. Snir, "The Performance of Multistage Interconnection Networks for Multiprocessors," *IEEE Transactions on Computers* **C-32**, No. 12, 1983.

[8] F. A. Tobagi, T. Kwok, and F. M. Chiussi, "Architecture, Performance, and Implementation of the Tandem Banyan Fast Packet Switch," *IEEE Journal on Selected Areas in Communications* **9**, No. 8, pp. 1173-1193, 1991.

[9] A. V. Krisnamoorthy and F. E. Kiamilev, "Fanout, Replication, and Buffer Sizing for a Class of Self-Routing Packet-Switched Multistage Photonic Switch Fabrics," Photonics in Switching Meeting, paper PThC4-1, March, 1995.

[10] T. J. Cloonan, "Comparative study of optical and electronic interconnection technologies for large asynchronous transfer mode packet switching applications," *Optical Engineering*, Vol. 33, No. 5, pp. 1512-1523, 1994.

[11] A. W. Lohmann et al., in *Digest of the Conference on Optical Computing*, (Optical Society of America), Washington, D. C., 1985, paper WA3.

[12] A. W. Lohmann, "What Classical Optics Can Do for the Digital Optical Computer," *Applied Optics*, Vol. **25**, pp. 1543-1549, 1986.

[13] G. Eichmann and Y. Li, "Compact Optical Generalized Perfect Shuffle," *Applied Optics*, Vol. 26, pp. 1167-1169, 1987.

[14] S.-H. Lin, T. F. Krile and J. F. Walkup, "2-D Optical Multistage Interconnection Networks," *Proc. SPIE*, vol. 752, pp.209-216, 1987.

[15] K.-H. Brenner and A. Huang, "Optical Implementations of the Perfect Shuffle Interconnection," *Applied Optics*, vol. 27, pp. 135-137, 1988.

[16] C. W. Stirk, R. A. Athale, and M. W. Haney, "Folded Perfect Shuffle Optical Processor," *Applied Optics*, Vol. **27**, pp. 202-203, 1988.

[17] A. A. Sawchuk, I. Glaser, "Geometries for Optical Implementations of the Perfect Shuffle," *Proc. SPIE*, Vol. 963, p. 270, 1988.

[18] M. W. Haney and J. J. Levy, "Optically Efficient Free-space Folded Perfect Shuffle Network," *Applied Optics*, Vol. **30**, No. 20, pp 2833-2840, 1991

[19] G. C. Marsden, P. J. Marchand, P. Harvey, and S. C. Esener, "Optical Transpose Interconnection System Architecture," *Optics Letters*, Vol. 18. No. 13, 1993.

[20] M. W. Haney, "Pipelined Optoelectronic Free-Space Permutation Network," Optics Letters, Vol. 17, No. 4, pp. 283-285, 1992.

[21] IEEE/LEOS Topical Meeting on Smart Pixels, IEEE Digest Catalog No. 94TH0606-4, July, 1994.

[22] M. Decina F. Masetti, A. Pattavina, and C. Sironi, "Shuffleout Architecture for ATM Switching," ISS'92 Vol. 2, IEICE, Oct., 1992.

[23] M. Schwartz, Telecommunications Networks: Protocols, Modeling, and Analysis, Addison-Wesley Publishing Company, Reading, MA, 1988.

[24] M. W. Haney, "Self-similar Grid Patterns in Free-space Shuffle/Exchange Networks," *Optics Letters*, Vol. 18. No. 23, pp. 2047-2049, 1993.

[25] M. W. Haney and M. P. Christensen, U. S. Patent # 5,467,211, "Optoelectronic Sliding Banyan Network," issued November 14, 1995.

[26] Haney, M. W. and Christensen, M. P., "Performance Analysis and Optical Interconnection Module Evaluation for the Free Space Sliding Banyan Network," Photonics Switching '96, April 24, 1996.

[27] R. R. Michael, M. P. Christensen, and M. W. Haney, "Experimental Evaluation of the 3-D Optical Shuffle Interconnection Module of the Sliding Banyan Architecture," *IEEE Journal of Lightwave Technology*, September 1996.

3.4 Exerimental Evaluation of of the 3-D Optical Shuffle Interconnection Module of the Sliding Banyan Architecture

3.4.1 Introduction

3.4.1.A Motivation

Free space optical interconnects (FSOI) offer the potential to overcome the performance limitations of conventional interconnection technologies in many parallel computing and network switching applications. The advantages of FSOI stem from combining the inherent plane-to-plane mapping ability of optical systems, where the interconnection medium is free space, with the emerging optoelectronic (OE) smart pixel technology, where the input and output (I/O) and computing functions are performed by integrated arrays of emitters, detectors, and logic. Interconnection systems that use the basic imaging property of lenses have a fundamental ability to pointwise interconnect many thousands of channels with no interchannel coupling (cross-talk) within the interconnection medium As an added benefit, free-space optical systems perform plane to plane interconnections *without physical connectors or wire pads,* creating a natural I/O density improvement over 1-D chip edge connected metallic technologies or even 2-D array technology. The advantages of 3-D FSOI architectures will be fully exploited only as smart pixel technology pushes toward reliable integration of arrays of surface normal emitters and detectors with VLSI densities of electronic logic. Fortunately, monolithic and hybrid smart pixel technologies are making rapid strides.

A benefit of using 3-D FSOI is the higher *aggregate bandwidth* provided by the high spatial parallelism of optics. This can be exploited by simply imaging an array of emitters directly onto an array of detectors in a straight pass-through interconnect. While this interconnection relies on the spatial distribution of high bandwidth transmitters for a large aggregate bandwidth, it does not take advantage of the optical elements to interchange the positions of data in the array. A measure of the amount of interchange in data needed for an interconnect is its minimum *bisection bandwidth* (BSBW). Bisection bandwidth is defined as the product of the bandwidth of a single link, B, and the number of links which would have to be removed to divide a network into two equal parts. For networks with N nodes, the BSBW can range from 0 for a straight pass-through interconnection (no data interchange) to $BN^2/2$ for a fully connected crossbar switch. In simplest terms, the BSBW is a direct measure of the parallel interconnection difficulty because it quantifies the number of "wire-crossings" necessary to achieve the desired link pattern. The effects of time skew are implicit in the definition of BSBW since high BSBW networks have links between nodes very close together and other links to nodes on the opposite sides of the array. The use of imaging optics, which tend to make all path lengths roughly equal, leads to an improvement in the skew of the system.

Architectures for switching, sorting, and distributed computing often use multi-stage interconnection networks (MINs) which require data from different, widely separated nodes as input for local computation. The perfect shuffle (PS) [1] and related link patterns commonly used in MINs have a BSBW of $BN/2$. Since the number of links connecting nodes in PS based systems with large distances across the array grows linearly with N, the link length will eventually limit the bandwidth of the electronic signal. Electronically interconnected MINs are faced with the conflicting requirements of high BSBW and large physical size (required by the large number of nodes). This limits the scalability of all–electronic switching approaches. No such intrinsic limit exists for an architecture utilizing optical interconnections, since the link bandwidth is independent of link length and link crossings for optical channels.

3.4.1.B Optical Sliding Banyan Architecture

Optical versions of the PS were proposed [2, 3] and several implementations of the optical PS and higher order shuffles have been demonstrated [4 - 10]. These approaches implement MINs by shuffle interconnecting sequential planes of smart pixels. An interleaved clustering approach was suggested [11] to take advantage of the very high intrinsic BSBW of the optical PS, the projected density of logic and optical I/O, and the identical interconnection patterns for every stage in a shuffle interconnected MIN. This concept uses a single OE backplane and a retro-reflective FSOI architecture to interconnect all stages of the MIN. With this approach, resources are distributed laterally across a single multichip backplane, rather than longitudinally across many smaller single chip planes. This single plane partitioning is more amenable to the use of conventional electronic packaging, such as multi-chip modules (MCMs). The single backplane, interconnected with an interleaved shuffle FSOI system is a central feature of the Sliding Banyan (SB) architecture [12,13].

The SB concept is shown in Figure 1. In practice, the SB will employ a single OE backplane on which all stages of the MIN are co-located and interleaved to implement an efficient deflection routing scheme [13,14]. For illustrative purposes Figure 1 displays the SB in an unfolded and noninterleaved manner. Data packets are destination-tag routed, i.e., switched at each stage on one bit of the destination address (1 switches the packet to the lower port, 0 switches the packet to the upper port). If a potential conflict occurs, the "losing" packet is switched incorrectly (dashed lines), and begins its routing anew at the next stage. Hence, a new "virtual" banyan has effectively "slid" to accommodate that packet's routing needs. In the actual SB architecture, all the switching planes are interleaved in the same physical plane – the OE backplane. This groups all stages for each node into a cluster (e.g., for 1024 node 50 stage SB there would be 1024 clusters each with 50 emitters and 50 detectors) allowing them to share a common exit point from the network. This greatly reduces the switching and output resources of the SB architecture, even for nonuniform traffic [14], and

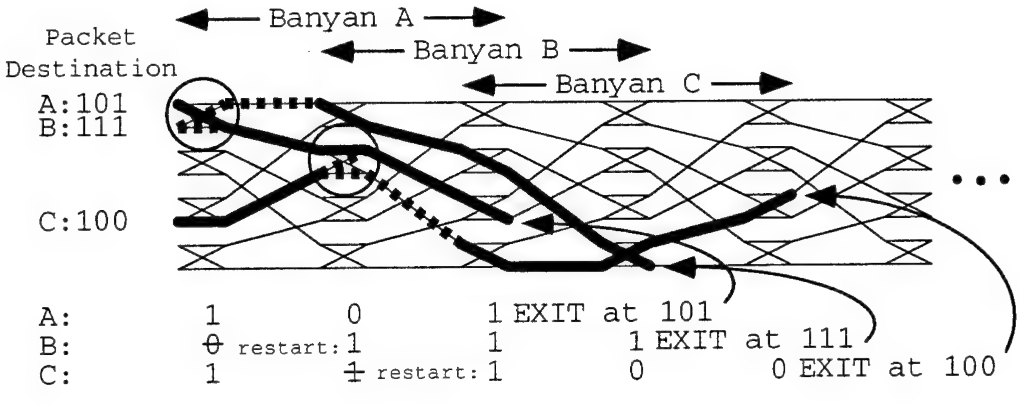


Figure 1. SB network depicting the shuffle MIN topology and the destination tag self-routing scheme. The routing of 3 packets (A, B, and C) with different destinations is shown. A is routed directly to its output location 101 where it is removed from the network (a 1 means exit at the lower port of the 2 x 2 switch and a 0 means exit at the lower port). A and B conflict at the first 2 x 2 switch and B is misrouted once and begins rerouting at the next stage, i.e., B's banyan is slid as shown (switch routing priority is determined by a priority tag). A and C collide at the second switch where C is misrouted and begins rerouting at the next stage. B and C exit at their destinations 111 and 100.

enhances the scalability of the SB for large networks. This results from the SB's novel resource partitioning scheme and efficient routing algorithm.

The SB requires a *single* optical module to implement many stages of parallel optical shuffle interconnection. To achieve this the I/O for the parallel links are grouped into clusters. The electronic logic for switching and removing packets from the network is located at this cluster, thereby providing distributed control of the switching network. This clustering 3-D topology has been shown to provide significant advantages in switching, control and interconnect resources [14]. The SB optical interconnection module must therefore pointwise interconnect the clusters of optical I/O across the entire OE backplane in the requisite shuffle link pattern.

In this paper the development and experimental evaluation of the SB's FSOI module is discussed. Its ability to achieve the requisite critical resolution and registration is evaluated. This work is the first experimental demonstration of a FSOI system that employs a single OE backplane (10 cm x 10 cm) that is interconnected with a macro-optical refractive lens array in a reflective architecture. The key elements and relevant system design issues are presented in the following sections. Section II describes the requirements for the lens array and the individual lenses that comprise the array. The layout of the simulated optical I/O on the OE backplane is also detailed in this section. Section III describes the experimental optical interconnection system and results. Section IV discusses the alignment procedures necessary to make the SB interconnection system practical. This approach reduces the degrees of freedom in the alignment problem, enabling the procedure to be easily automated. Section V highlights key issues related to the implementation of the optical SB architecture.

3.4.2 Optical Interconnection Requirements of the SB Architecture

All of the interstage interconnection patterns for the SB architecture are identical, making it possible to interleave the interstage optical I/O on a single OE backplane and implement the interstage FSOI with a single macro-optical refractive lens array. A schematic diagram of this SB interconnection scheme is shown in Figure 2. The OE I/O is laid out so that each node in the backplane is a cluster of emitters and detectors as shown. Within each cluster, the layout for the emitters is identical to that for the detectors. The emitters and detector clusters may be physically separated or interleaved, depending on the smart pixel circuit requirements. The clusters of OE I/O are distributed across several optoelectronic integrated circuits (OEICs) which comprise the OE backplane. For example, Figure 2 shows 16 clusters per OEIC. The number of OEICs in a given SB switch will be determined by the amount of real-estate available in a chip and the maximum allowable heat dissipation within a chip. Optical I/O densities of 100s per cm^2 appear to be within the reach of VCSEL based smart pixel developers [15]. Therefore, a 256 node, 25 stage SB would require 16, approximately 2 cm x 2 cm OEICs. The layout of the OE backplane, i.e., the chip placement and the I/O locations, would be determined by smart pixel circuitry requirements and optical and electronic packaging constraints. Optically efficient and MCM compatible grid patterns for placement of I/O on a given OEIC and the placement of the OEICs in the backplane has been developed based on the self-similarity properties of fractal grids [16]. Such a chip and I/O cluster pattern is depicted in Figure 2.

The identical shuffle interconnection pattern between each stage in the SB is implemented with a single 2-D array of lenses and a planar mirror that folds the optical signals back onto the backplane. In this architecture, the OE I/O is partitioned so there is one lens for each OEIC in the OE backplane. Each lens in the array acts as both a transmitting element and a receiving element. The overall effect of the lens system is to reproduce a unity magnification image of the input emitter cluster at its destination detector cluster on the backplane. The basic shuffle system is shown unfolded in Figure 3-a. Light from each emitter of each cluster (emitter clusters are shown as squares) is collimated and directed bny the transmitting lens to a receiving lens that focuses the light onto the associated detector in the detector cluster (detector clusters are shown as triangles). Figure 3-b shows how in a reflective system light from two emitters in

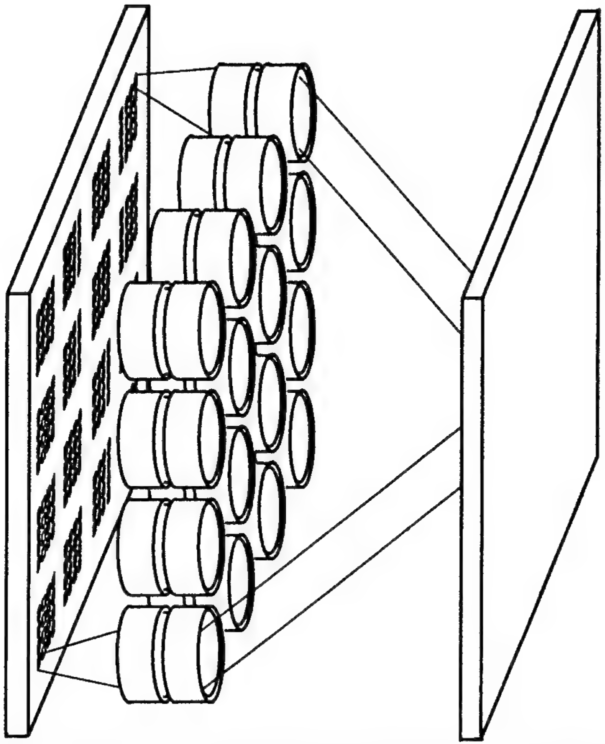


Figure 2. Schematic diagram of the SB optical interconnection system showing the layout of the I/O on the backplane, the shuffle optics which is comprised of a refractive lens array and a planar mirror. The reflective architecture allows the optical I/O for all stages of the SB to be co-located on a single OE backplane.

a cluster is reflected by a planar mirror and is re-focused by the corresponding receiving lens onto the appropriate detectors within a detector cluster. Note that the emitter and detector clusters (shown as squares and triangles, respectively) are depicted side-by-side in Figure 3-b. In practice it may better to interleave the two clusters. The solid lines in Figure 3 indicate the mapping of emitter clusters to the corresponding detector clusters.

The critical implementation issues for the SB FSOI module are resolution (i.e., blurring of imaged emitters) registration (i.e., errors in mapping emitters onto detectors) and the optomechanical alignment needed in the interconnection module Typical element sizes for the individual emitters (VCSELs) and detectors within the clusters are envisioned to be ~ 10 μm and ~ 30 μm, respectively. Since the folded and interleaved SB optical system images each emitter onto a corresponding detector in a one-to-one shuffle mapping operation, the registration accuracy must be on the order of 10s of μm *across the entire multichip array.* Furthermore, the individual emitter blur spot size due to diffraction and aberrations in the lens array must also be ~ 10s of μm so that a significant fraction on the emitter's energy is captured by the associated detector and crosstalk is minimized. Current photolithographic techniques will provide sub-micron placement accuracy of the OE I/O within the individual smart pixel chips and chip "pick-and-place" equipment can achieve on substrate accuracy of ~ 10 μm across an MCM. These are well within the accuracy required for the OE backplane. Therefore, the main implementation issues in the SB stem from the imaging capabilities of the FSOI module.

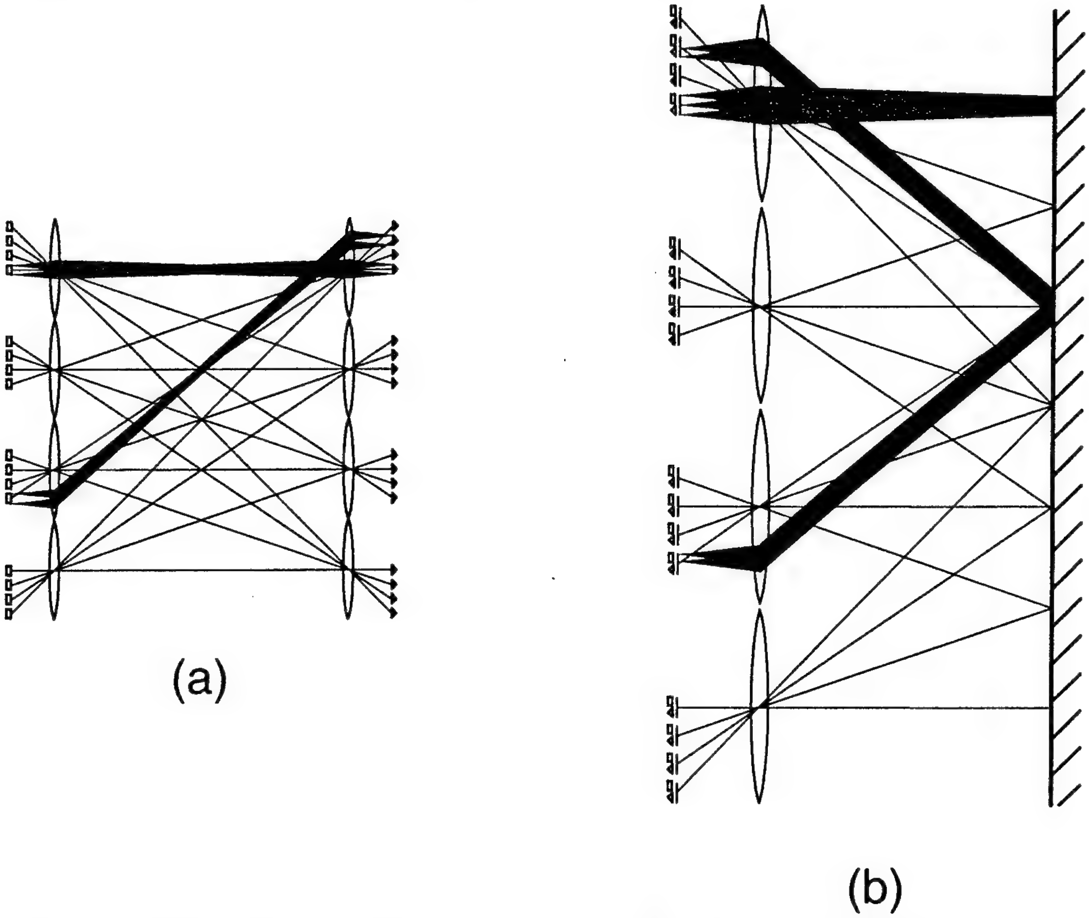


Figure 3. 3-a shows a schematic diagram of the unfolded SB shuffle optics. The squares represent emitter clusters and the triangles represent detector clusters. The "beams" show the mapping of 2 individual emitters within the cluster onto corresponding detectors. 3-b shows a schematic of the folded optics. All of the optical I/O is interleaved on a single backplane.

3.4.2.A Lenses and the 2-D Macro Lens Array

Several key requirements of the lenses and the lens array are highlighted in Figure 3. First, the lenses must have a wide field of view (FOV) and a low f/# in order to yield a compact optical system. Second, the image produced by the transmitter/receiver pair must be nearly free from aberrations in order for each emitter in a cluster to register properly on the corresponding detector. Aberrations will lead to blurring of the individual emitter images and improper registration of the emitters clusters on the detector clusters. Third, the lenses in the array should have a large aperture to mounting barrel diameter ratio (ideally 1) so that the lens elements themselves are in close proximity to one another. This will allow for optimal placement of the OE I/O and will yield the smallest dimensions for the OE backplane as shown in Section II-B. Finally, each lens in the array should be nearly identical in focal length and optomechanical packaging to assure accurate mapping of the emitters onto the corresponding detectors on the OE backplane.

Eventually the SB FSOI module will employ custom optical elements. However, the initial experimental module described in this paper uses off–the–shelf lenses. Applications such as high performance cameras, laser scanning systems, and projection systems require lenses with performance characteristics similar to those of the lenses in the SB FSOI module. Experimental evaluation of several candidate lenses, using a linear 1-D and 2-D array of VCSELs, has been conducted [13]. Of all the lenses tested, the best candidate for the SB interconnection module is a projection lens with an f/# of 1.12, an effective focal length of 13 mm and a relatively large aperture to mounting barrel diameter ratio of 0.72.

The lens array in the SB FSOI module required 16 nearly identical projection lenses. Therefore, a large number of these lenses were purchased and experimentally evaluated. The initial set of lenses were all well matched optically (i.e., effective focal lengths all matched to ± 0.5%). However, experiments indicated that the working focal lengths, the distance from the lens mounting barrel to the lens focal plane varied by more than a few percent. Since the individual lenses in the lens array were not individually focused, the working focal lengths of the lenses had to be experimentally evaluated and carefully matched. These evaluations yielded a set of nearly identical projection lenses, both in terms of effective focal lengths and working focal lengths, that were incorporated into the SB interconnection module.

3.4.2.B Optical I/O Layout

The parameters of the selected lenses for the macro-lens array constrain the placement of the optical I/O in the simulated OE backplane of the SB FSOI module. In the OE I/O design process there are three key constraints that must be considered. They stem from simple geometrical arguments and the physical limit on heat dissipation within an OEIC. The design constraints relate key lens parameters to possible OE I/O spacing and cluster spacing in the SB FSOI module.

The first constraint on the placement of optical I/O in the backplane stems from the active lens aperture of the transmitting lens. The OEICs in the corner of the chip array will have OE I/O sites farthest off-axis of their transmitting lens. The I/O site farthest from this lens's center must be within the active lens aperture. In fact, the assumed square array of I/O must be completely circumscribed by the circular lens aperture as shown in Figure 4-a.

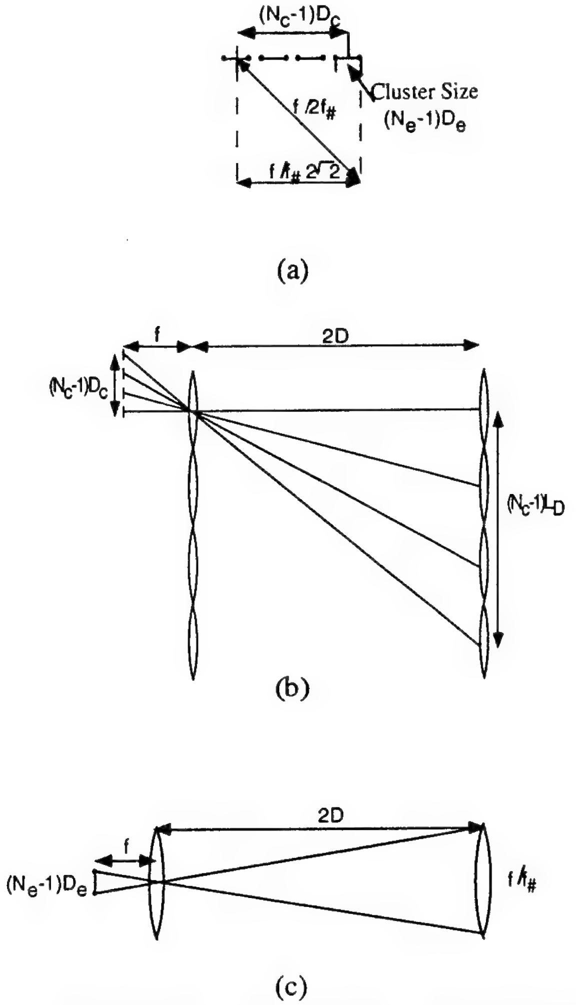


Figure 4. 4-a shows a top view of the lens and I/O geometry used in deriving the upper bound on the cluster spacing as a function of emitter spacing. 4-b and 4-c depict the geometry used to derive the lower bound on cluster spacing as a function of emitter spacing. 4-b shows that the light emitted by the individual clusters must be captured by their respective receiving lenses, which cannot overlap, and 4-c depicts the fact that the transmitted cluster must fall within the active aperture of its receiving lens.

The second constraint insures that all emitters in a single cluster are captured by the *receiving* lens and that the receiving lens array spacing does not require the physical overlapping of lenses. The third constraint is specified as the minimum spacing between I/O sites. This would come directly from the physical limitations, e.g., power density limitations on the emitter spacing. For the prototype design a typical minimum spacing of 100 μm is used.

Figure 4-a depicts the first geometrical constraint. The farthest off-axis emitter of the transmitting lens must fall within the active aperture of this lens. Therefore, distance to the edge of the active lens aperture, ($f/2f_\#$) must be greater than the position of the farthest emitter from the lens axis:

$$\frac{f}{2\sqrt{2}f_\#} \geq D_c(N_c - 1) + \frac{D_e(N_e - 1)}{2}, \tag{1}$$

where f is the focal length of the lens, D_c is the distance between cluster centers, N_c is the number of clusters behind one lens (i.e., on one chip) along one dimension, D_e is the distance between emitters (I/O sites) and N_e is the number of emitters in one dimension of a cluster. This leads to an *upper* bound on the cluster spacing as a function of emitter spacing:

$$D_c \leq \frac{\left(\frac{f}{2\sqrt{2}f_\#} - \frac{D_e(N_e - 1)}{2}\right)}{(N_c - 1)}. \tag{2}$$

A *lower* bound on the cluster spacing as a function of emitter spacing is found by considering the cluster spacing required to insure that the receiving lens array has no physical overlap of lenses. This places a physical lower bound on the cluster spacing as a function of the distance to the mirror, D, or equivalently the distance to the receiving array in an unfolded architecture. This relation depicted in Figure 4-b, is:

$$D_c \geq \frac{fL_D}{2D}, \tag{3}$$

where L_D is the diameter of the lens barrel and D is the distance from the lens plane to the mirror in the retro-reflective architecture. This is a lower bound since the cluster spacing, D_c, can always be increased from this point. This simply separates the lenses, allowing some space between them. This equation relates the cluster spacing to the minimum system length. As the cluster spacing increases the minimum system length decreases. This is intuitively correct, as the increased cluster spacing creates larger angles in the system, thereby shortening it. This can only be exploited until the upper bound is reached. At this point the numerical aperture of the lenses is fully utilized.

To relate Equation 3 to emitter spacing requires a calculation of the maximum emitter spacing as a function of the distance to the mirror, D. This relation is depicted in Figure 4-c. The cluster size, $D_e(N_e$-1) must be small enough to be captured by the active aperture of the receiving lens, f/f#. Equation 4 is the relation of distance to the mirror, D, to the cluster spacing, D_e:

$$\frac{D_e(N_e-1)}{f} \leq \frac{f}{2D\sqrt{2}f_{\#}}. \qquad (4)$$

Equations 3 and 4 combine to form a lower bound on the cluster spacing as a function of emitter spacing:

$$D_c \geq \frac{\sqrt{2}(N_e-1)D_e}{\beta}, \qquad (5)$$

where β is the ratio of the lens barrel diameter to active lens aperture. This is a line with slope proportional to 1/β. Points above and to the left satisfy this constraint, while smaller cluster spacing and larger emitter spacing do not. Values of β approaching 1 are advantageous as they allow more freedom within the design constraints. This is intuitively correct, as a lens with its active aperture equal to its physical dimensions (β=1) would be ideal. Equation 5 explicitly demonstrates the self similar grid requirements [16] of the optical I/O arrangement and the lens array placement. The limit on the portion of the OE backplane used for optical I/O is a scaled version of the useable portion of the lens array.

A final constraint is the physical limit on how small the emitter spacing can be. Figure 5 shows an example of all the design constraints for the experimental SB optical interconnection module. The vertical line at 100 μm is a lower limit on emitter spacing, the nearly horizontal upper line, Equation 2, is an upper bound and the sloped line, Equation 5, represents a lower bound. The area inside the triangular region formed by the three constraints contains all possible combinations of emitter spacing and cluster spacing for the specified lens parameters. The length of the optical interconnection module monotonically decreases with an increase in cluster spacing. So the minimum volume system would be designed to a point in the upper left corner of the design triangle in Figure 5. The design should be near this corner while avoiding both bounds, as the upper bound vignettes about half of the light of the outermost emitter at the transmitting lens and the lower bound vignettes at the receiving lens.

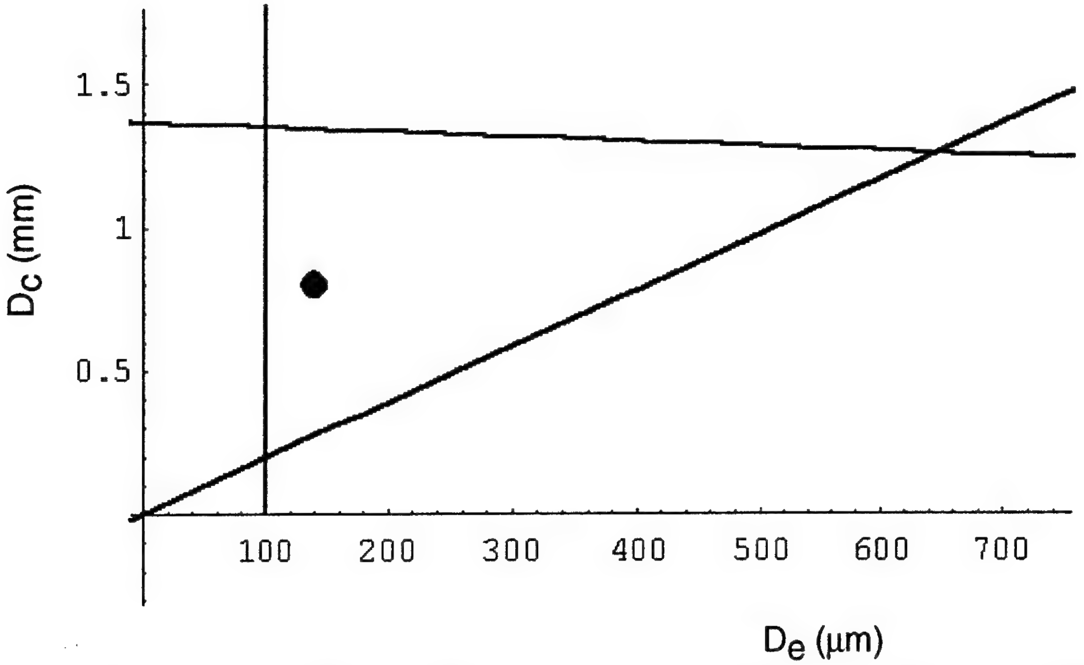


Figure 5. Example of the optical I/O layout design constraints for the SB interconnection module. The area inside the triangle represents all possible combinations of emitter spacing and cluster spacing for the experimental SB interconnection module. The dot shows the emitter spacing and cluster spacing used in the experiments.

In the simulated OE backplane used in the experimental SB interconnection module, a 100 µm minimum between emitters was specified. The point at 140 µm emitter spacing and 800 µm cluster spacing represents the prototype developed. This design point, shown in Figure 5, was chosen as it was near this upper left hand corner of the design triangle. This point provided a good design buffer from both vignetting constraints under the assumption of fairly small (< 10 degrees is typical of VCSELs elements characterized for this setup in our laboratory) emitter beam divergence. The physical minimum bound on emitter spacing used in the prototype is representative of VCSEL based smart pixel technology. In a final system this limit would be approached as it is constrained by the smart pixel fabrication process and does not affect the alignment and optical efficiency (i. e., vignetting) of the FSOI module.

3.4.3 Interconnection Evaluation:

3.4.3.A *Experiment and Results*

A. Test Module

Using the lenses and the OE I/O layout described above, an experimental optical module for the SB architecture was constructed in the laboratory. The system was comprised of 256 clusters (nodes) each with 5 I/O sites. This led to a module with 1280 shuffle interconnected links. The overall dimensions of the system is approximately 10 cm x 10 cm x 20 cm.

A schematic diagram of the experimental set up used to evaluate the SB FSOI module is shown in Figure 6. The experimental system is comprised of three planes as illustrated. The bottom plane holds the simulated OE backplane, the middle plane is used to mount the lenses that make up the lens array and the top plane is used to hold the planar front surface reflecting mirror. These planes are supported so that the inter-plane distance and parallelism may be adjusted. In the experiments, the backplane was used as the reference plane to which all other elements in the system were aligned.

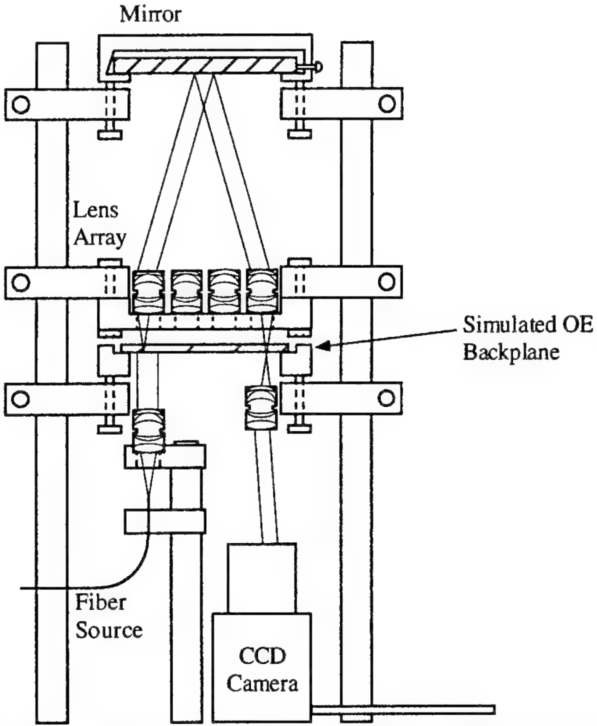


Figure 6. Schematic diagram of the experimental set up used to evaluate the SB 3-D FSOI module.

The layout of the simulated OE backplane is depicted in Figure 7. Figure 7-a shows the layout of the I/O sites on the entire backplane. The sites are arranged in 16 groups which contain 16 clusters of 5 I/O sites. Each one of the groups represents a single OEIC. Figure 7-b shows an expanded view of a single group (i.e. for a single OEIC), and Figure 7-c depicts the pattern for a single cluster. This backplane I/O pattern was fabricated on a chrome-on-glass plate using a photolithographic process and etching process similar to that used in semiconductor device manufacturing. The chrome background on the plate is 10 % transmissive and the etched circles representing the emitter and detector sites are transparent. This partially transparent design enabled the interconnect resolution and registration accuracy to be evaluated as discussed in detail in Section III-C. The diameter of the holes simulating emitters and detectors is 25 μm. The square clusters measure 140 μm on a side and the spacing between clusters is 800 μm.

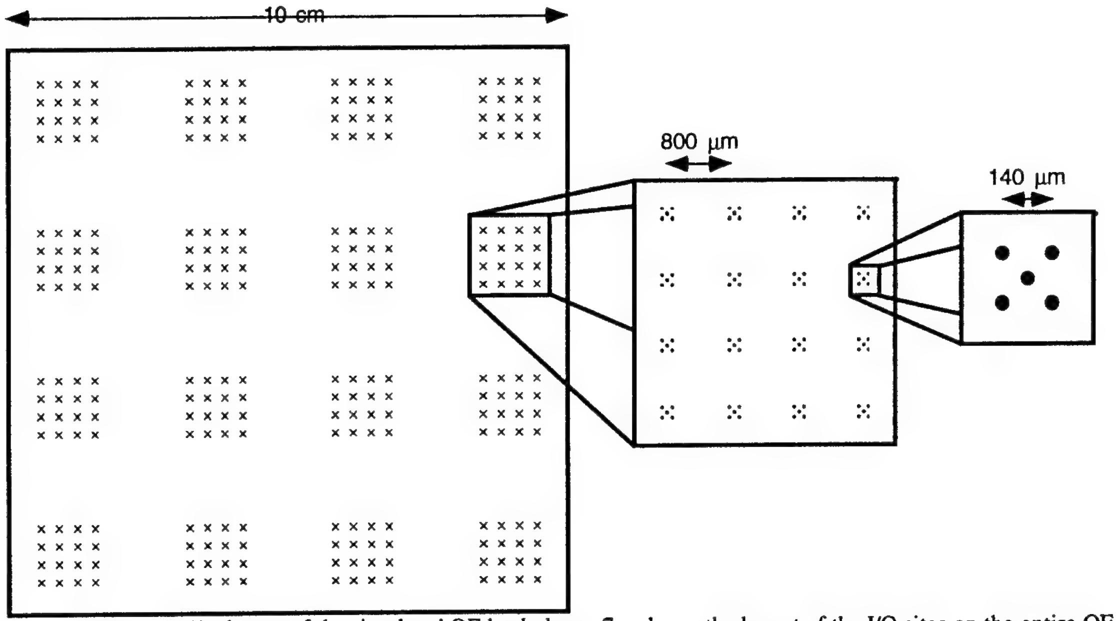


Figure 7. The layout of the simulated OE backplane. 7-a shows the layout of the I/O sites on the entire OE backplane. The sites are arranged in 16 groups which contain 16 clusters of 5 I/O sites. Each one of the groups represents a single OEIC. 7-b shows an expanded view of a single group (OEIC), and 7-c depicts the pattern for a single cluster.

The lens array plane is comprised of a flat aluminum plate with 16 apertures, one for each lens in the lens array. A photograph of the top view (mirror removed) of the lens array and fixture is shown in Figure 8. Three lenses have been removed to shows the chrome-on-glass mask containing the OE I/O sites in the simulated backplane. A lens was precisely aligned above each group of I/O sites (OEIC) using the self-alignment procedure discussed in the next section.

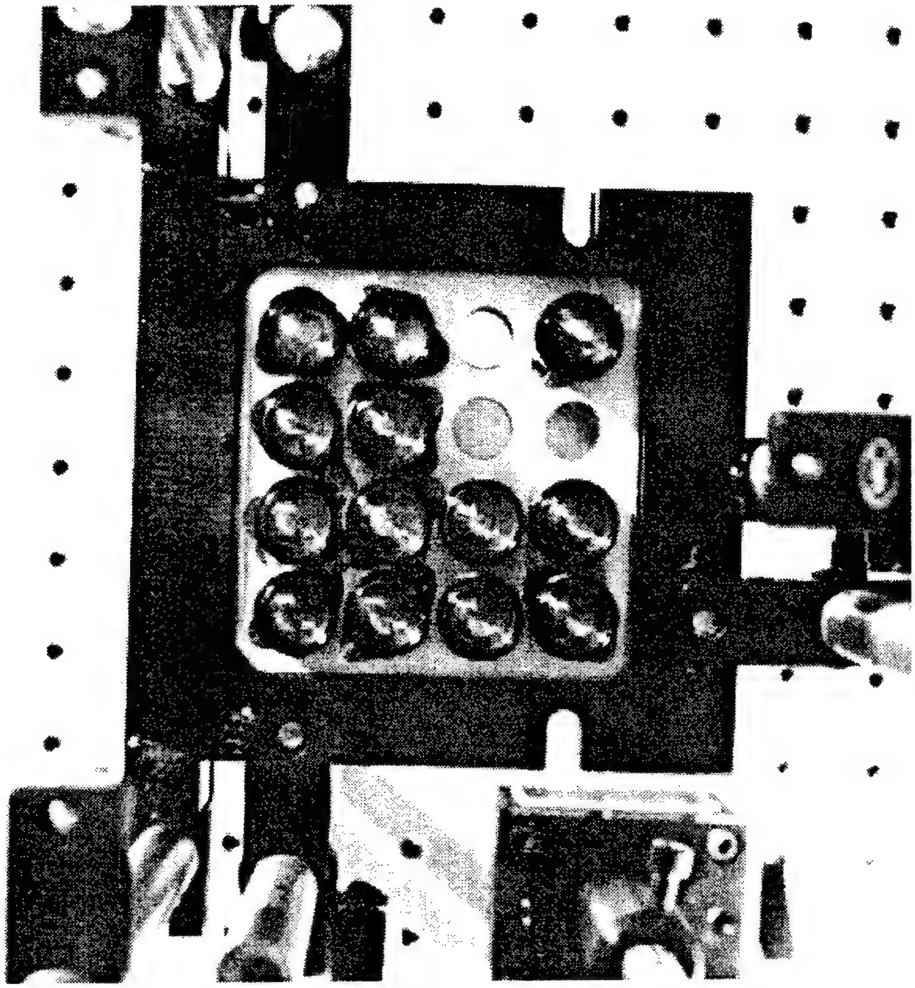


Figure 8. A photograph of the top view (mirror removed) of the lens array and fixture. Three lenses have been removed to show the mask containing the OE I/O sites in the simulated backplane below the lens plane.

B. Alignment of FSOI Module

A simplified alignment procedure was developed for the experimental FSOI module. The OE backplane was used as the reference to which all other elements were aligned. The distance and interplane parallelism between the lens array fixture and the backplane were iteratively adjusted until the backplane was located in the focal plane of each lens in the lens array. This was accomplished by moving a lens to different positions in the array plane and verifying that the incident beam was focused on the backplane. Next, the approximate height of the mirror was set so that the diagonal corner to corner interconnection was possible, i.e., all emitters in one corner cluster in the OE backplane could be captured by the corresponding receiving lens and imaged onto the corresponding detector cluster. Then the mirror plane was adjusted until it was parallel to the OE backplane. The exact mirror height and parallelism are not critical since none of the lenses in the lens array are fixed at this point. Any error in mirror height and mirror tilt will be compensated for during the positioning of the lenses. At this point in the alignment procedure, the three planes, OE backplane, lens plane, and mirror plane, were set parallel to one another with the proper interplane distances.

Next, the lenses in the lens array were individually placed and aligned on the lens fixture. The alignment of the lenses was accomplished using a self-alignment technique. Each lens was positioned in x and y (on the lens fixture plane) so that its on-axis focal point registered with the central I/O site of the corresponding cluster in the simulated OEIC under the lens. This enabled the lenses to be positioned to within a few µm of their prescribed location in the lens array. The interlens alignment relied upon the self-alignment of all of the lenses in the array. In other words, once all the individual lenses were self-aligned, the entire optical interconnection system was aligned. The simplicity of this approach makes it amenable to automation.

C. Evaluation

The SB FSOI module was evaluated by illuminating a cluster from underneath the simulated OE backplane with light from a HeNe laser ($\lambda = 632.8$ nm) that was guided into the module with a 1 mm plastic optical fiber and was nearly collimated with a miniature projection lens as shown in Figure 6. The light that passes through the simulated emitter cluster (holes in the chrome mask) is collimated and directed by the transmitting lens, is reflected by the planar mirror, and is imaged onto the corresponding simulated detector clusters (holes in the chrome mask) by the corresponding receiving lens in the lens array. The chrome background in the mask was 10 % transmissive so the exact location of the focused spots at that “targeted” detector cluster could be observed through the mask. This made it possible to compare the location of the detector sites to the location of the corresponding imaged emitter sites that are transmitted through the interconnection module.

The location of the detector sites and the relative position and size of the transmitted images of the emitter sites were evaluated using an imaging system comprised of a CCD camera, an image relay lens, and video lens that is placed under a receiving lens as shown in Figure 6. A representative image obtained using this system is shown in Figure 9. The clusters in this image represents interconnects that were made when light was transmitted from six other emitter clusters at different OEIC locations on the OE backplane. To measure the misalignment between the detector sites and imaged emitter sites (i. e., the registration error) or blurring of the imaged emitters (i.e., resolution) each detector cluster was enlarged (as shown in the insert in Figure 9) and carefully analyzed. Due to the magnification in the image relay optics each pixel in the insert in Figure 9 represents approximately a 6.5 μm square in the OE backplane, so for the interconnection in the enlargement (which shows the worst case registration error) the registration accuracy is ~ 20 μm and the resolution is ~ 10 μm. By moving the HeNe illumination system and CCD imaging system to various locations in the simulated OE backplane all possible interconnections can be evaluated. Carrying out this process shows that the worst case registration accuracy of this system is ~ 20 μm. However, when the 4 most closely matched lenses are used to test various interconnections this registration accuracy is improved to ~ 10 μm (< 1.5 pixels) for all possible interconnects. The spot sizes were nearly diffraction limited images of the object emitters, given the narrow beam divergence of the sources. The beam diameter at the lens surface was approximately 1 mm, leading to a diffraction blur size of ~8 μm. This corresponds well with the measured ~35 μm image spot sizes (which is roughly the convolution of the ~8 μm blur with the ~25 μm hole size in the mask, thus the optical system is nearly diffraction limited. This resolution and registration accuracy was verified for all possible interconnections in this simulated OE backplane.

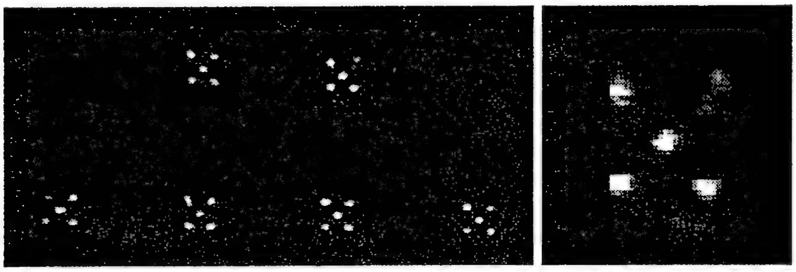


Figure 9. Image showing the simulated emitter clusters that have been transmitted from 6 other OEIC sites on the backplane. The magnified image shows the worst case registration error. The bright spots result when the transmitted emitter beams hit the prescribed detectors sites. The faint double image results when the emitter misses the detector site. In this case the background light transmitted through the detector site and the imaged emitter site that shows through the 10 % transmissive background appear to have the same intensity.

3.4.4 Discussion

A systematic alignment procedure for an optical interconnection module relies on controlling and decreasing the degrees of freedom (DF) present in the system. The alignment procedure developed for the SB FSOI module has been developed so that at most 2 DF need to be addressed simultaneously. A logical choice for an alignment reference in the SB FSOI module is the OE backplane. Once the backplane is fixed, there are 8 reduced DF that need to be addressed in this alignment procedure: 3 each for the lens plane and mirror plane, and 2 for the position of individual lenses within the lens plane. For the mirror and lens plane the 3 DF correspond to the distance to the backplane and 2 angles of tilt. For the individual lenses the 2 DF are the x and y placement in the lens array plane. Note that the lens placement is restricted to the lens plane, in other words, no individual focusing or tilting of the lenses in the lens array is required in this procedure.

The 8 DF are reduced to 5 after fixing the lens plane parallel and at a focal length distance from the OE backplane. This procedure, while iterative, progresses rapidly and is easily automatable. The mirror plane is fixed using a procedure similar to that used for the lens plane. The placement of the mirror plane reduced the DF to the 2 needed for lens positioning. Neither the mirror nor the lens array plane needs any rotational alignment. All rotational alignment issues are addressed in the x - y self alignment of individual lenses within the lens array. This elimination of rotational alignment is an important feature of the reflective SB interconnect architecture. Another benefit of the lens self-alignment procedure is that only rough mirror alignment is necessary as each lens in the interconnection module would be aligned to compensate for a slight misalignment of the mirror.

Self-alignment is the key to the success of this procedure and is facilitated by the reflective architecture. In order to execute this procedure, a detector site is placed where the on-axis focal point of each lens should register. A lens is moved in the x and y directions within the lens array plane until its on-axis focal point rnegisters with the self-alignment detector site. In this way, all of the lenses in the array are self-aligned. The self-aligned lens, acts as the emitting and receiving lens, thereby collapsing the 4 DF present if there were 2 separate lenses, into 2 DF. The optical interconnect is designed so that when all lenses are self aligned the global interconnection pattern is completed. This reduction of the DF to 2, is only possible in a retro-reflective architecture. This single OE backplane system removes rotational DF that would be present between each 2 planes of a multiplane architecture, as well as the mutual alignment of separate sending and receiving lenses. This procedure minimizes the DF necessary to align the FSOI in the SB system and facilitates an automated alignment procedure, which is critical for the manufacturing of optoelectronic systems.

The single plane reflective architecture is relatively insensitive to the exact mirror plane height and tilt and completely eliminates any rotational alignment tolerances in packaging the FSOI module. also allieviates structural rotation between planes in a multi-plane packaged FSOI module. Therefore, this single plane system minimizes the number of compononets which require critical alignment and stabilzation. This single plane approach allows the utilization of current pick-and-place technology for backplane population, and therefore is compatable with current MCM packaging systems. The mirror plane is invariant to rotation, and has very little depenecy on its distance from the backplane.

3.4.5 Conclusions

The FSOI approach used in the SB module has been experimentally evaluated. The experimental results yielded ~10 μm resolution and registration accuracy for all possible interconnections across a 10 cm x 10 cm simulated OE backplane when lenses that were carefully chosen to be nearly optically and optomechanically identical were used in the macro-lens array. This is on the order that is needed for this architecture since 10 μm emitters and 30 μm detectors are envisioned as smart pixel technology evolves. If the emitters are mapped to within 10 μm of their prescribed locations with minimal blurring then a large fraction of the transmitted energy from each emitter will be captured by the corresponding detector and the interconnection will be completed. As expected, the experiments showed that the registration accuracy for a given interconnection depends strongly on the lenses used to make that particular interconnect. Small differences in the optical or optomechanical parameters (i.e., effective focal length, position of focus relative to lens mounting barrel, etc.) of the lenses in the lens array leads to larger registration errors. Therefore, in order to implement a SB FSOI module the lenses that make up the lens array must be very well matched. In the future this will be accomplished by using custom lenses that have very tight optical specifications and optomechanical tolerances.

The retro-reflective FSOI SB architecture is shown to have several important advantages over multi-plane FSOI architectures in both optomechanical packaging as discussed in this paper and performance [13, 14]. Only a single OE backplane, a single macro-lens array, a planar mirror, and a relatively simple optomechanical package are required to implement all stages of an SB MIN. The optomechanical alignment of the system is easily automatable since at most two variables need to be adjusted at one time. This is facilitated by the simple lens self-alignment technique that has been developed to align the lenses within the lens array. Once all the lenses are individually self-aligned all of the lenses in the array are automatically aligned. By locating all of the optical I/O on one backplane all and employing this alignment technique, all rotational alignment and lens to lens alignment issues are eliminated. Finally, since each lens in the system acts as both a receiving and transmitting lens there is a natural symmetry in the optical system that can be exploited to minimize aberrations in the macro-lens array.

The results of this experimental work suggest that the reflective 3-D free space shuffle interconnection architecture when used with emitter based smart pixel OE technology could implement a 1024 node network, operating at 1 Gigabit/sec/channel.

This suggests that an SB based ATM switching fabric can be scaled to aggregate throughputs >1 Terabits/sec.

3.4.6 References

[1] H. S. Stone, "Parallel Processing with the Perfect Shuffle," *IEEE Trans. on Computing*, vol. **C-20**, pp. 81-89, 1971.

[2] A. W. Lohmann, et al., in *Digest of the Conference on Optical Computing*, (Optical Society of America), Washington, D. C., paper WA3, 1985.

[3] A. W. Lohmann, "What Classical Optics Can Do for the Digital Optical Computer," *Applied Optics*, vol. **25**, pp. 1543-1549, 1986.

[4] G. Eichmann, and Y. Li, "Compact Optical Generalized Perfect Shuffle," *Applied Optics*, Vol. **26**, pp. 1167-1169, April 1987.

[5] S.-H. Lin, T. F. Krile and J. F. Walkup, "2-D Optical Multistage Interconnection Networks," *Proc. SPIE*, vol. **752**, pp.209-216, 1987.

[6] K.-H. Brenner and A. Huang, "Optical Implementations of the Perfect Shuffle Interconnection," *Applied Optics*, vol. **27**, pp. 135-137, Jan. 1988.

[7] C. W. Stirk, R. A. Athale, and M. W. Haney, "Folded Perfect Shuffle Optical Processor," *Applied Optics*, vol. **27**, pp. 202-203, 1988.

[8] A. A. Sawchuk and I. Glaser, "Geometries for Optical Implementations of the Perfect Shuffle," *Proc. SPIE*, vol. **963**, p. 270, 1988.

[9] M. W. Haney and J. J. Levy, "Optically Efficient Free-space Folded Perfect Shuffle Network," *Applied Optics*, vol. **30**, No. 20, pp. 2833-2840, July 1991.

[10] G. C. Marsden, P. J. Marchand, P. Harvey, and S. C. Esener, "Optical Transpose Interconnection System Architecture," *Optics Letters*, vol. 18, pp. 1083-1085, July 1993.

[11] M. W. Haney, "Pipelined Optoelectronic Free-Space Permutation Network," *Optics Letters*, vol. **17**, pp. 283-285, Feb. 1992.

[12] M. W. Haney and M. P. Christensen, "Optical Freespace Sliding Tandem Banyan Architecture for Self-Routing Switching Networks," International Conference on Optical Computing, Aug. 1994.

[13] M. W. Haney and M. P. Christensen, "Sliding Banyan Network," *Journal of Lightwave Tech.*, May 1996.

[14] M. W. Haney and M. P. Christensen, "Sliding Banyan Network Performance Analysis," submitted to *Applied Optics*, Jan. 1996.

[15] T. Nakahara, S. Matsuo, S. Fukushima, and T. Kurokawa, "Performance comparison between multiple-quantum-well modulator-based and vertical-cavity-surface-emitting laser-based smart pixels," *Applied Optics*, February, 1996.

[16] M. W. Haney, "Self-Similar Grid Patterns in Free-Space Shuffle/Exchange Networks," *Optics Letters*, vol. **18**, pp. 2047-2049, Dec. 1993.

3.5 Topological Aspects of Free-space Optical Interconnections

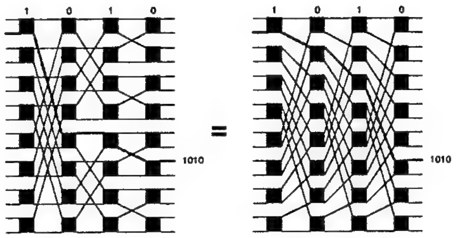
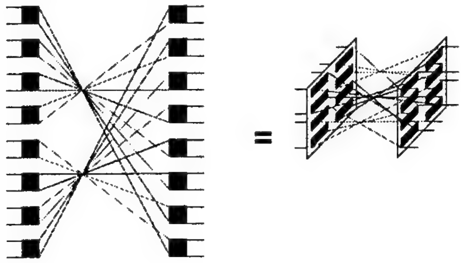
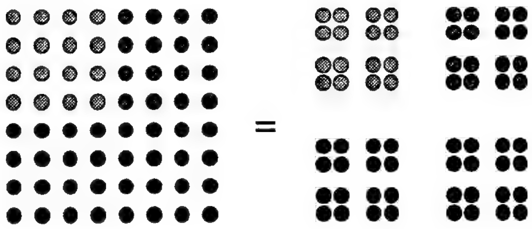
Systems that use smart pixel-based free-space optical interconnects (FSOI) provide two general capabilities for overcoming interconnection bottlenecks. We refer to the first as "intelligent parallel data *transfer*," and the second as "intelligent parallel data *interchange*." Optical imaging provides a high throughput approach to linking smart pixel planes for data transfer. For data interchange, FSOI provides the means to perform the partitioning and interleaving needed to implement PS-like link interconnection patterns. Several 2-D optical PS approaches have been demonstrated. Both the data transfer and interchange capabilities exploit the 2-D high density I/O capabilities of smart pixels. Our focus is on the use of FSOI for *data interchange.* In particular we address implementation issues for Multistage Interconnection Networks (MINs) that are based on the banyan network – a well-known point-to-point building block for switch architectures that can use packet self-routing.

The application of FSOI techniques to a network architecture can be viewed as a mapping of the network's functional interconnection pattern onto a 3-D optical interconnection architecture. Such a mapping amounts to a topological transformation which preserves the interconnections and functionality of the switch configuration, but achieves performance advantages owing to the use of 3-D space and smart pixel capabilities. In fact the architecture can be represented as a series of topological transformations that each exploit a performance advantage of photonic interconnects. The cumulative performance advantage of a FSOI implementation of a switch architecture derives from the aggregate advantages of several distinct geometrical transformations of the link interconnection pattern.

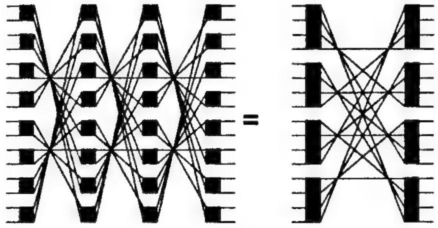
Examples of topological transformations that apply to the FSOI banyan-based networks and the motivation for using them are shown in the following table. When taken together, in a multistage switch configuration the retro-reflective interleaved architecture provides clustering of switching, interconnection, and control resources. The clustering provides the means to reduce the redundancy in resources required by conventional interconnection networks.

An example of enhancements made possible through the combination of the above outlined topological transformations are outlined in Table 5 for a Sliding Banyan Architecture. Since the architectures in Table 5 are topologically equivalent, the switching resources are identical. However, since the SB architecture is within a factor of 2-3 of the minimum for switching resources [1], no other architecture will gain an advantage over the FSOI SB. *Control and signal I/O are the biggest issues in high throughput switching* and dominate the resource requirements. As shown in the table, the photonic SB topology provides *fundamental* improvements in scaleability in output and control resources. The table shows significant reductions in the most troublesome resource elements, such as chip output drivers and control, while eliminating thousands of co-axial cables. This is truly significant since control resources are the dominant element of large switch designs.

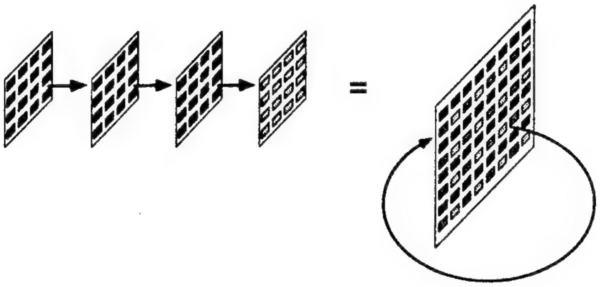
Table 5. Topological Transformations of FSOI Networks used in the GMU concept.

Description	**Schematic**
1. Using a shuffle link pattern, that is isomorphic to the butterfly pattern, between stages of the banyan – to simplify the optical design	
2. Formatting the shuffle as a 2-D shuffle, rather than a 1-D shuffle – to take better advantage of optical and MCM packaging techniques.	
3. Arraying the smart pixel on self-similar grids, rather than rectilinear grids – to increase multichip pixel density and optical efficiency.	

4. Using shuffles of order higher than 2 (e.g., 4), to reduce latency and take advantage of VLSI high local complexity.



5. Spatially interleaving multiple stages – to cluster nodes and thereby reduce the amount of required electronic routing resources in the smart pixel.



6. Using a retro-reflective approach – to distribute the smart pixels across a single backplane, simplify optical alignment, and reduce the number of output drivers required.

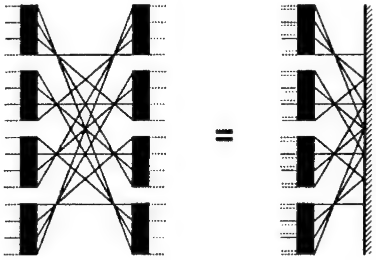


Table 6. Comparison of Sliding Banyan Resource Scaleability for 10^{-6} Blocking Rate

Resource Type	Photonic SB Requirement	Electronic SB Requirements	Comment (N=1024)
Switching: Internal Electronic	$5(2\times2)(N/2)Log_2N$	$5(2\times2)(N/2)Log_2N$	same
Switching: Output Electronic	$4NLog_2N$	$4NLog_2N$	same
Chip Output Drivers	N	$4NLog_2N$	×40 in power
Interconnect Drivers	$5NLog_2N$ VCSEL / detector pairs.	$5NLog_2N$ high speed interchip line drivers	×5-×10 in power
Interconnection Medium	~16 lenses, mirror, and free-space	$5NLog_2N$ coax cables	50,000 coax cables are impractical.
Control	$\propto N$	$\propto 5NLog_2N$	×50 in gates

Some of the transformations' performance improvements apply to the FSOI implementation issues, such as electronic interface and packaging, optical alignment, optical efficiency, and pixel I/O density. Other transformations have a significant impact on the total amount and complexity of control, I/O, and interconnection resources required to achieve a given low blocking rate traffic pattern. One of the most significant performance enhancements stems from the topological transformations that spatially clusters nodes in a common backplane. It is shown that, for self-routing banyan-based networks that use a deflection routing scheme, the overall control resources are reduced by a factor of 1/M, where M is the number of stages in the network. M can be a fairly large number; e.g., $M \cong 50$ for $N = 1024$ in some deflection routing schemes which demand low blocking rates. This reduction is a significant improvement over other implementations since control resources tend to dominate the requirements of high throughput networks. Furthermore, with clustering, the number of output drivers needed at the interface to the fabric is equal to N. This is a significant reduction from the worst case number of output drivers (N×M) for non-clustered implementations. These control and output resource reductions stem directly from the 3-D topological transformations that, when applied to the network, exploit optics' inherent ability to efficiently interchange data. The resulting resource efficiency provides a significant improvement in the size, weight, and power needed to achieve a given performance specification. The performance enhancements achieved by these topological transformations are made practical *only* through the use of 3-D FSOI. High aggregate throughput switching is thus a good candidate application for smart pixel-based FSOI architectures.

3.6 Fundemental Geometric Performance Advantages of Free-space Optical Interconnections

3.6.1 Introduction / Motivation

With smart pixel throughput capabilities projected to exceed 1 Tbit/s/cm^2 [1], the use of free-space optical interconnects (FSOI) may provide significant advantages over all-electronic interconnection technologies. The combination of parallel high density I/O with local electronic logic in smart pixels enables new architectures for a large class of interconnection-limited problems. With the rapidly emerging high density capabilities of smart pixels, a systematic approach for quantifying the advantages of FSOI is needed.

Systems that use smart pixel-based FSOI provide two general capabilities for overcoming interconnection bottlenecks: "intelligent parallel data *transfer*," and "intelligent parallel data *interchange*." Optical imaging provides a high throughput approach to linking smart pixel planes for data transfer. In this case the high I/O density of smart pixels may provide a power consumption and size advantage over electronics. For data interchange, FSOI provides the *additional* ability to perform the data partitioning and interleaving useful in space variant link interconnection patterns like the perfect shuffle (PS) [2], which are inherently difficult to implement in planar interconnection technologies. Such patterns are characterized by high *bisection bandwidth* (BB) [3]. The BB of a network is defined as the bandwidth that crosses a boundary that cuts the network in half – it is a measure of wiring difficulty. In architecture design, there is a direct trade-off between minimum BB and latency in a network. It is therefore generally desirable to implement networks with the largest minimum BB that can be practically achieved to solve a given problem. The ability of optical elements to interconnect large arrays in space-variant patterns, without crosstalk in the medium, suggests that FSOI techniques are particularly promising for problems with high BB. This paper outlines a general model for quantifying the fundamental interconnection advantages of FSOI over planar electronic interconnections in architectures requiring a large level of data interchange.

3.6.2 Approach

The baseline used to compare FSOI and electrical interconnection technologies is the total circuit substrate area, A_T, required to achieve the desired interconnection network [4]. A_T provides a good basis for comparison because it is readily applied to estimations of volume, latency, and weight. Furthermore, for substrate regions in which the bandwidth capacity is fully utilized, and in which the circuit electrical interconnection elements are considered to be lumped capacitive loads – such as within an MCM, the total circuit interconnection power will scale linearly with area. The total circuit interconnection area is thus an implicit measure of the size, weight, and power consumption associated with implementing an interconnection network.

To compare the circuit areas required to implement an architecture with optical interconnections and metallic interconnections, the architecture's interconnection complexity must first be examined. The total aggregate bandwidth, B_A, of any architecture, i.e., the product of the number of links and the bandwidth per link, can be divided into two portions: bandwidth within the network and bandwidth into or out of the network, as illustrated in Figure 1a. Any given implementation's interconnection capacity is therefore determined by the chosen technology's capacity for internal interconnections, i.e., within a VLSI chip, and the external bandwidth capacity to any other interconnection layer, i.e., from the chip to an multi-chip module (MCM) or printed circuit board (PCB).

To determine the interconnection requirements within a given technology for any network, the network is analytically parsed as follows. The network is first divided into equal sized sub-networks. Figure 1b shows an arbitrarily selected bisection of the network in Figure 1a. The partitions each have smaller internal bandwidth requirements than the original network, at the cost of additional inter-partition, or external, bandwidth. The partition boundary should be chosen so that it minimizes the inter-partition bandwidth created between any two subnetworks. For example, if the network is divided into two subnetworks, then the partition boundary is selected to be along the network's minimum bisection division as shown in Figure 1c. By definition this places the minimum bandwidth between the partitions and divides the network in half. Each resulting subnetwork is similarly parsed leading to a tree representation for the parsing analysis. (In general the bisection may not be the optimal partitioning scheme and partitions of other sizes should considered. However, most interesting network topologies, such as cross-bars and multistage interconnection networks (MINs), will use bisection partitioning due to their symmetry. The analysis in this paper is therefore restricted to bisection partitioning. However the basic approach generalizes to higher order k-sections (where k is prime.)) Figure 2 depicts such a tree. Each node of the tree is labeled with that partition's internal BB and its external bandwidth requirements.

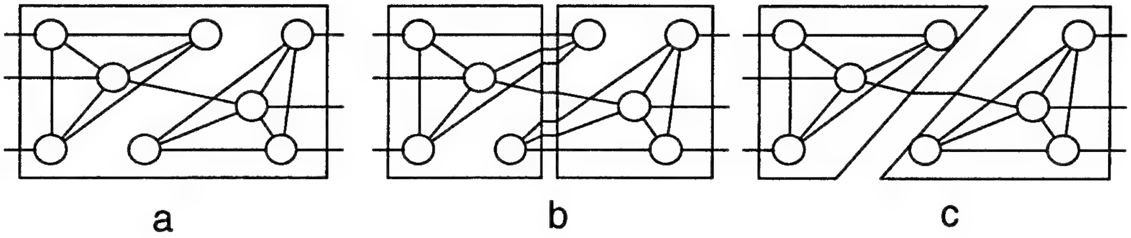


Figure 1. Network representations. a: arbitrary interconnection network, b: network arbitrarily divided into two equal subnetworks, c- network optimally divided into 2 equal subproblems – the minimum bisection partition.

In order to make a valid comparison between an optical implementation of an architecture and an electrical interconnection approach, the optimum partitioning of the network must be found for each technology and the resulting systems compared. An initial lower bound on electrical substrate area is determined for the best initial partitioning of the architecture into 2 equal parts. For electrical interconnections, if this minimum bisection requires total bandwidth B to be routed in the substrate, then the size requirements can be examined for the system by considering the maximum bandwidth density B_{max} (bandwidth/cm) in each substrate. Therefore, the inter-partition bandwidth requirements of the network can be directly translated into partition substrate area requirements [5]. The width, W, of a square substrate is equal to the inter-partition bandwidth divided by the bandwidth density of the given technology. If the required partition width W_{Chip} ($B/B_{max,CHIP}$) exceeds the maximum size of a single chip, then MCM-level interconnections must be utilized. Furthermore, if the maximum MCM size is exceeded by W_{MCM}, then PCB-level interconnections must be utilized. Since the bandwidth density of PCBs is less than that of MCMs, which in turn is less than VLSI on-chip densities, the substrate area required is greatly impacted by requiring that interconnections be placed into the lower density substrate levels.

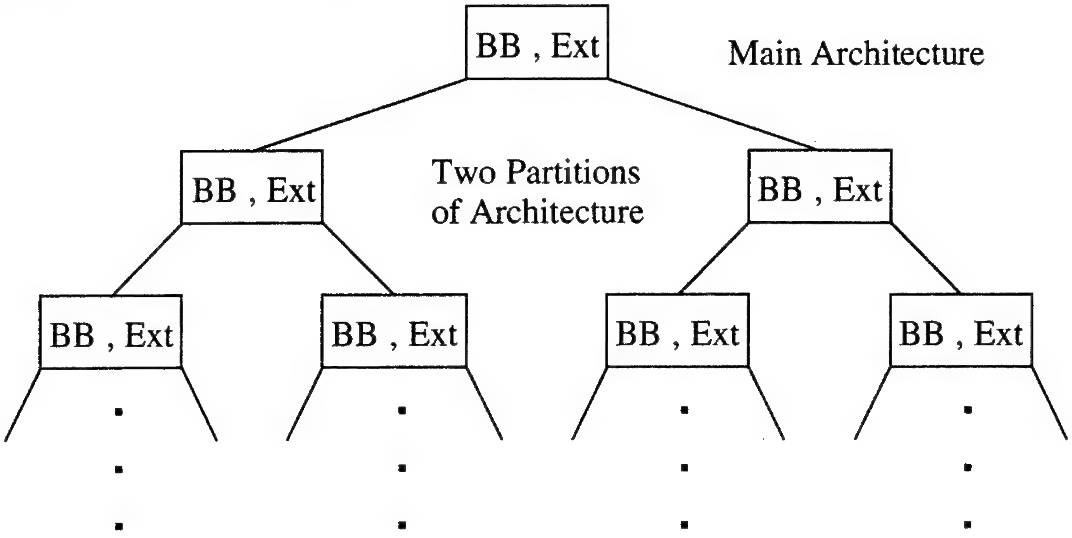


Figure 2. An example bisection tree used to create sub-networks for interconnection area requirements analysis.

A "globally" interconnected network architecture is characterized as having its greatest bisection bandwidth partition at its first division. By this definition, the lower bound for globally interconnected architectures is tight, i.e., the substrate area required for the network, as calculated after the first bisection partitioning, is the maximum substrate area that will be needed for interconnection purposes. It is possible, however, that the first bisection of an architecture does not present the greatest interconnection challenge because the overall network is not globally connected, e.g., each pair of stages in a MIN is globally interconnected, but the entire network is not. If this is the case, then the lower bound determined by the first bisection is not tight and therefore does not provide the best estimate of the actual area required to implement the architecture. In this case the substrate area predicted is necessary but not sufficient for implementing the network. The bound can be improved by examining each partition created in the first bisection. One of the two partitions may be a globally interconnected sub-network (e.g., as in a pair of stages in a MIN), causing the next bisection to determine the substrate area requirements. By repeatedly examining the results of bisections and determining how each affects the substrate area, the lower bound can be tightened. When all sub-partitions have been repeatedly bisected until no processors remain connected, the lower bound on area is tight and the area is sufficient.

To evaluate area requirements for metallic interconnection package technologies, the inter-partition and intra-partition bandwidth capacity of each possible routing layer must be determined. This is derived from the throughput density of each of the packaging technologies. The Chip-MCM-PCB packaging hierarchy represents current state-of-the-art electrical interconnection technology. It is assumed that each level of packaging is square in shape and is populated with a regular grid pattern that evenly distributes the processing nodes, and their I/O, across the substrates. Under these assumptions, Figure 3 depicts the limit of internal bandwidth and inter-partition (external) bandwidth for planar electrical interconnections. The maximum electrical BB for such circuit packages is determined by the total bandwidth that may cross a boundary that cuts the number of nodes into 2 equal halves. To achieve high BB performance, this boundary should be as large as possible. It would be tempting to draw, therefore, a meandering boundary that is longer than any dimension of the substrate. However, this would result in a BB density that exceeds the density of the technology for the next partitioning into smaller subnetworks, as described above. The internal bisection bandwidth is limited, therefore, to the product of the *width* of the physical partition (as shown in the figure) and the throughput density of the partition's technology.

The inter-partition bandwidth is limited by product of the length of the partition perimeter, as shown in Figure 3, and the throughput density of the technology responsible for the inter-partition connections – all inter-partition interconnects must cross the perimeter of the partition while in the lower layer. This is true even if the actual interconnections are distributed across the area of the substrate, as, for example, is the case in flip-chip packaging technologies. Typical maximum sizes for the substrates of these elements are W_{Chip}=2.5, W_{MCM}=15, and W_{PCB}= 45 cm. Typical maximum bandwidth densities for the three substrate technologies are B_{Chip}=0.5, B_{MCM}=0.2, and B_{PCB}=0.05 Tbit/s/cm. The maximum internal and external interconnection bandwidths for the chip are therefore estimated to be $B_{int,Chip} = B_{Chip}W_{Chip}$ = 1.25 Tbit/s and $B_{ext,Chip} = 4B_{MCM}W_{Chip}$ = 2 Tbit/s. The maximum internal and external interconnection bandwidths for the MCM are similarly estimated to be $B_{int,MCM} = B_{MCM}W_{MCM} = 3$ Tbit/s and $B_{ext,MCM} = 4B_{PCB}W_{MCM} = 3$ Tbit/s. If the required bandwidth capacity of a partition is greater than the above stated capacity of a chip, then the interconnection must take place in an MCM layer. Similarly, if the MCM bandwidth capacity is exceeded, then the routing must take place in the PCB. Once it is decided in which layer the routing takes place for the inter-partition bandwidth, the required substrate size is determined. The substrate cross-section required equals the required bandwidth B divided by the routing layer technology's bandwidth density. For instance, if a partition has a BB of 2 Terabits/sec, then it cannot be routed within a chip – it must be routed in an MCM substrate, and, since B_{MCM}=0.2, it therefore requires a 10 cm × 10 cm MCM substrate.

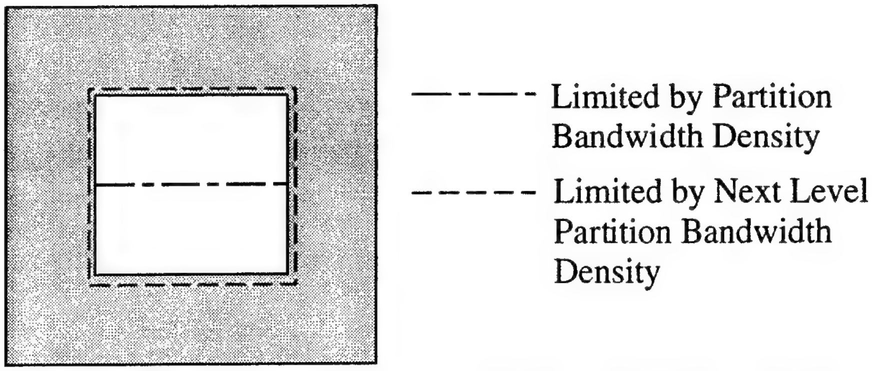

Limited by Partition Bandwidth Density
Limited by Next Level Partition Bandwidth Density

Figure 3. Internal and external interconnection bandwidth partition boundaries for metallic interconnection technologies.

Since FSOI interconnections are not confined to planar links, the interconnection density limitations stem from the *area* I/O density capabilities of smart pixel technology and the ability of optical elements to perform the inter-chip data interchange functions. A large set of interesting interconnection patterns that can be performed by FSOI have been proposed and demonstrated. FSOI implementations of 2-D PS-based concepts for MINs [6-10] are examples of networks which are composed of globally interconnected subnetworks, i.e., pairs of stages. Other interleaved optical concepts require a single global network [11-13]. Other examples of globally interconnected FSOI networks include the cross-bar [14]. These concepts are all implementable with macro-lenslet arrays which perform the required data interchange and transfer across many nodes. Figure 4 depicts the basic interconnection bandwidth capabilities of smart pixel-based FSOI. As shown, simple macro-lenslet arrays can be used to point-wise link smart pixels across several chips, that are either located along the same plane (if a mirror is used) [11,12], or on facing planes. Each ray in the figure indicates a single point-to-point link in a high bisection width pattern, such as a cross-bar or shuffle. The bandwidth limitation is determined by the ability of the imaging lens elements to connect smart pixel chips with the high spatial density needed to fully exploit the smart pixel I/O density. FSOI concepts based on interleaved imaging of sub-arrays are able to link arrays of smart pixel I/O with resolution well beyond that required to achieve the anticipated Tbit/sec/cm^2 I/O densities of smart pixel arrays. The maximum bandwidth crossing external bandwidth boundaries for FSOI are therefore determined by the area of the smart pixel surface and the density B_D (Terabit/s/cm^2) of the optical I/O. In the figure, the external bandwidth boundary is indicated by the dashed line that is a side view of an area boundary equal to the area of the smart pixel array. If N chips are interconnected with a symmetric global data interchange patterns (e.g., a shuffle link pattern or crossbar) with I/O densities indicated as in the figure, then N/2 of the area I/O boundaries will bisect the nodes into 2 equal partitions. The BB capability of FSOI is thus given by 1/2 the total smart pixel I/O density for N chips. For example, if B_D =1 Tbit/sec/cm^2, then the bisection bandwidth capability of FSOI is $D_B A_c/2$, where A_c is the total smart pixel chip area employed. Unlike the metallic interconnection case, the FSOI BB capabilities are the same for internal (within a chip) and external (inter-chip) links within the network; i.e., there are no fundamental limit to total I/O density as the network bandwidth gets larger (such as exists in metallic interconnects when going from chip to MCM or from MCM to PCB.).

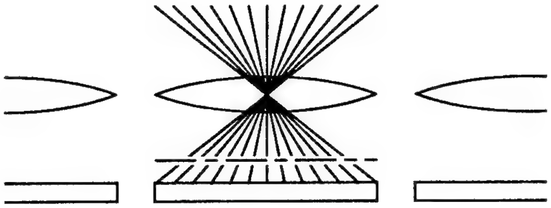


Figure 4. Schematic depiction of multichip FSOI, based on macro lenses, showing that the interconnection external bandwidth capability is determined by the I/O density of the smart pixel. The external bandwidth boundary is a plane above the smart pixel I/O (depicted from the side as the dashed line).

Given the internal and external bandwidth limitations for metallic interconnection technologies – as estimated above, the determination of physical partitions of the processing nodes represented in the bisection tree proceeds as follows. The bisection tree is climbed from bottom to top (as oriented in Figure 2), searching for the maximum partition size able to be implemented on a single chip. This is repeated for each of the lowest branches of the tree, until all nodes (leaves of the tree) have been placed into a "chip" partition. Any node is able to be placed into a chip partition if its BB is less than the bisection bandwidth of a chip and its inter-partition bandwidth is less than a single chip's interchip partition bandwidth as determined in the previous paragraph. Also, the children nodes must have BBs which are able to be implemented inside each half, i.e., they must have BBs requiring no more than 1/2 the chip diameter for routing. When these criteria are met, the parent node of all nodes in a chip partition dictates the size of the chip, as it must handle all of the bisection bandwidth and all of the off-chip bandwidth. Once the chip internal or external bandwidth density is saturated, subsequent partitions must be based on MCMs as the partition tree is climbed. This proceeds until the top node is reached or the MCM capabilities are saturated, at which point PCB partitions must be employed. After all chip, MCM, and PCB partitions have been defined for the given network, the total substrate area required for metallic interconnection technology may be determined by either the internal bisection bandwidth or external inter-partition bandwidth, whichever requires more substrate area. In general, the total metallic interconnection substrate area will be the total chip area, total MCM area, or total PCB area – whichever is largest.

Figure 5 gives examples of sub-network substrate areas required to implement *internal* BB at the various levels of metallic interconnection packages, based on the previously given estimates of maximum size and interconnection densities achievable on chips, MCMs, and PCBs. Since the BB capability is proportional to the width of the substrate (or $(\text{area})^{1/2}$ for square substrates) in metallic interconnections, the substrate area required is proportional to $(\text{BB})^2$ [4]. However, since the density of BB capability goes down as we transition from chip to MCM, and then from MCM to PCB, the substrate area required to implement a given BB is markedly increased, as shown by the steps in the figure. On the other hand, the area requirement for the optically interconnected networks does not suffer the step increases in area requirements, leading to a dramatic difference in substrate area required for FSOI over metallic interconnections. Furthermore, since the FSOI area requirement grows in direct proportion to the BB requirement [5], the slope of the FSOI line in Figure 5 is half that of the metallic interconnection lines. Similar arguments apply to determining the interconnection area as determined by *external* BB area limitations. Which metallic interconnection limit will dominate will be determined by the specific network inter- and intra- partition bandwidths and the specific substrate sizes and bandwidth densities of the metallic interconnection technologies being evaluated. Ultimately, the relative advantages of FSOI will be determined by whether internal or external bandwidth density constraints require the larger substrate area. In either case, however, Figure 5 depicts the important relative trends for metallic and FSOI interconnection technologies. It is seen that, for the selected smart pixel I/O capability and metallic interconnection capabilities given in the figure caption, the area requirement for metallic interconnections and FSOI are very similar for networks that can fit within a chip. However, if the BB is high enough to require the use of an MCM, the relative differences in substrate area become pronounced (greater than an order of magnitude). If the interconnection requirements exceed those of the MCM, then the greatest benefit of FSOI is obtained. In this regime, the difference in substrate area between metallic (PCB) and smart pixel ICs exceeds 2 orders of magnitude.

The analysis of required circuit area can be extended directly to system interconnection volume requirements based on reasonable assumptions of electronic packaging and projected performance of the optical system. The volume of electronic interconnection system is assumed to be directly proportionally to the total circuit area. We assume that, due to packaging and heat removal considerations, PCBs will have some minimum spacing (i.e., 2 cm). Similarly, MCMs will have some maximum packaging density (i.e., one half centimeter spacing) when they are not mounted on a PCB. VLSI Chips are assumed to be limited to stacking densities of ~10 per cm. These assumptions result in Figure 6, which shows the volume requirements for an architecture of a given BB.

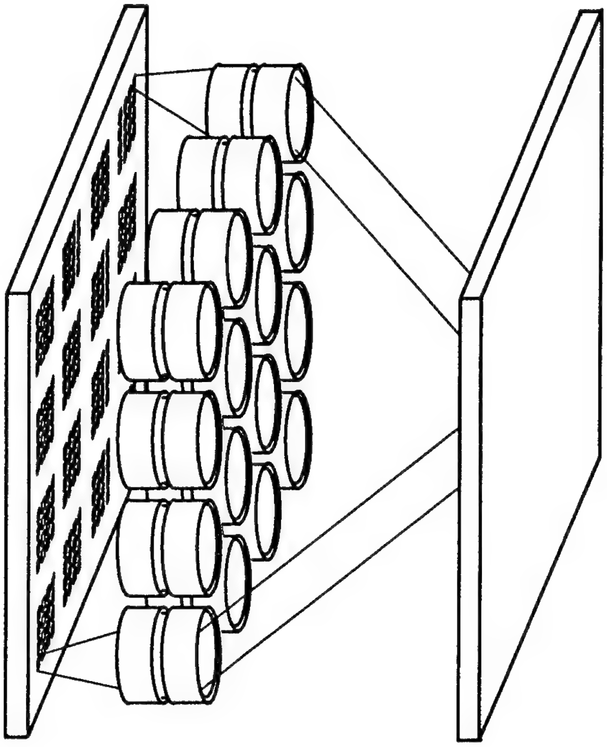


Figure 6. Multi-chip retro-reflective shuffle interconnection. Assuming f/1 optics, the modules shape approximates a cube.

An example retro-reflective optical system with f/1 optics is depicted in Figure 6 [13]. The total interconnection volume for this system is approximately a cube with sides equal to Area$^{1/2}$. The volume requirement can be further reduced by a factor of 2 by tilting the mirror to 45 degrees, and using an additional mirror to further fold the optical paths.

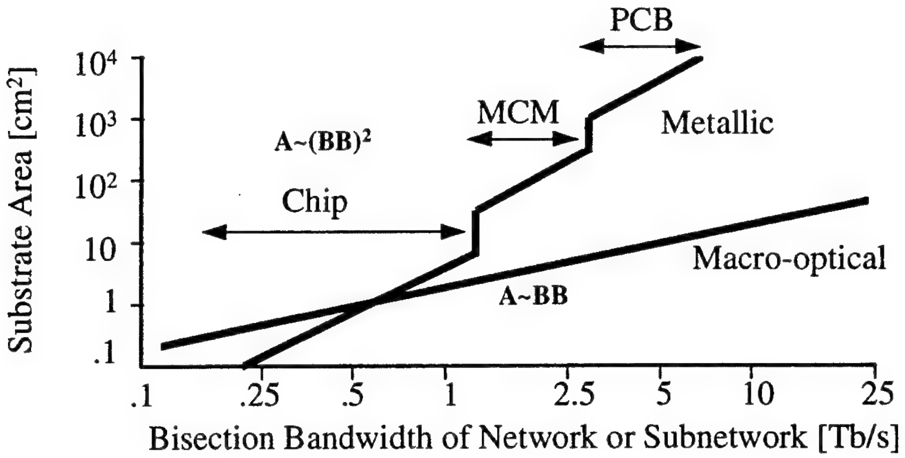


Figure 5. *Comparison of circuit area requirements for metallic and free-space optical interconnections. Maximum chip, MCM, and PCB interconnection densities are 0.5, 0.2, and 0.1, Tbit/s/cm, respectively. Maximum chip size is 2.5 cm. Maximum MCM size is 15 cm. Smart pixel I/O density is 1 Tbit/s/cm^2.*

The required volume for the macro optical system is therefore $V \cong \text{Area}^{3/2}/2$. This is plotted in figure 7. The volume requirement slope for the optical, as shown in the figure, is therefore less than that for the electrical case and indicates ~2 orders of magnitude advantage in volume for architectures requiring ~10 Tb/s and greater BB.

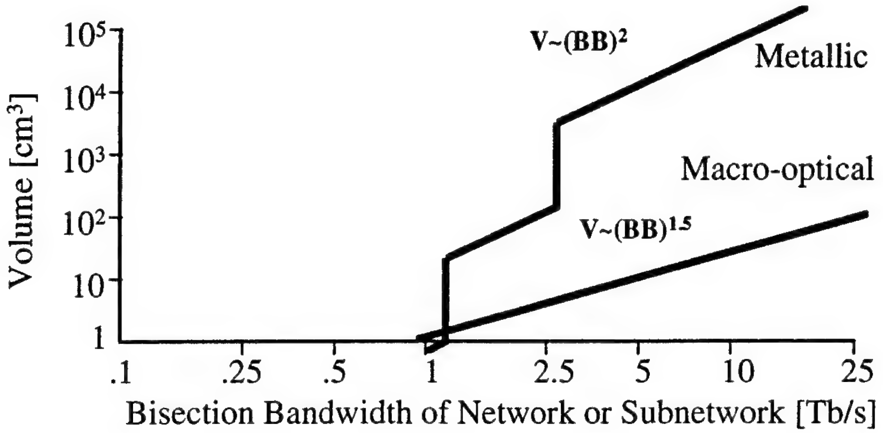


Figure 7. Comparison of metallic and optical interconnects on the basis of volume required.

To the extent that latency is directly proportional to the maximum path length in an architecture, these geometrical arguments can be extended to compare latency between metallic and optical networks. Such a comparison is shown in Figure 8. In both cases, the maximum path length is proportional to the width of the substrate, i.e., the square root of the area. Hence, for the electrical case, the maximum path length is proportional to the BB, whereas for the macro-optical case the maximum path length grows only as the square root of the BB. As shown in the figure, the latency advantage of FSOI is estimated to be two orders of magnitude for architectures with BBs of ~10Tbit/s or greater.

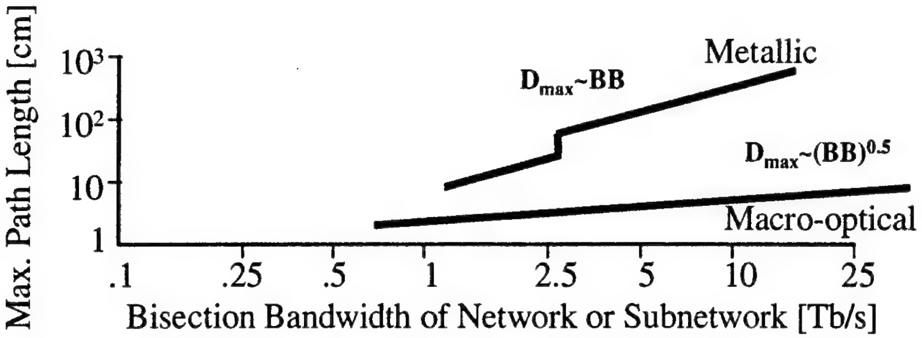


Figure 8. Comparison of metallic and optical interconnects on the basis of maximum signal pathlength

3.6.3 Discussion / Conclusion

It is clear from the data of Figure 5,6 and 8 that the benefits of FSOI become most pronounced when the network is globally interconnected and that *several* smart pixel chips are employed with maximum I/O density – for it is in this regime that metallic interconnection approaches exceed the capabilities of chips and MCMs and must rely on the lowest density technology (PCB) to achieve the required interconnections. For example, an array of 4 smart pixel chips, with a total area of ~ 8 cm^2 would achieve an interconnection BB of ~4 Tbit/s. The equivalent metallic interconnected network would require ~1000 cm^2 of substrate area – resulting in an approximately 2 orders of magnitude penalty in size.

A generic example of a high BB network that will be in this regime is an N×N cross-bar network that is characterized by $BB_{cross\text{-}bar} \approx BN^2/2$, where B is the bandwidth per link. Assuming the data of Figure 5, Table 1 compares the metallic and FSOI substrate areas required for implementations of cross-bars with various values of N. B is assumed to be 500 MHz. The table shows the very strong dependence of substrate area on N for the electrically interconnected cross-bar, while the optically interconnected offers a slower growth in substrate size with N. For N=256, it is noted that, for metallic interconnections, this corresponds to a PCB that is 1.6×1.6 meters in size! In this case it must be assumed that a bank of smaller boards would be used and that the inter-board interconnection density is maintained at the inter-board level by use of high-performance board-to-board interconnections of some kind. Clearly, this further compounds the problems of the metallic interconnection as compared with the potential compact embodiment of the FSOI cross-bar when implemented with high throughput smart pixel arrays.

Table 1. Comparison of Metallic and Free-space Optical Interconnection Substrate Areas for N×N Cross-bar Networks.

Number of nodes, N	FSOI cross-bar substrate area (cm^2)	Electrically interconnected cross-bar substrate area (cm^2)
64	2	4
128	8	1680
256	32	26880

Examples of other globally interconnected architectures include networks such as the smart pixel based Viterbi Decoder [11] and Sliding Banyan (SB) Switch [12-13]. Both of these architectures use spatially interleaved FSOI shuffle stages implemented as a single shuffle stage with multiple parallel channels. This approach yields topological advantages that reduce the redundancy in smart pixel control resources by physically co-locating multiple stages' smart pixel I/O into a single processing node. This results in globally interconnected architectures which grow with N according to the bisection bandwidth of a retro-reflective optical shuffle. In this case BB is approximately equal to BMN/2, where M is the number of parallel stages implemented in the interleaved shuffle architecture. For the Viterbi decoder M=4. For the SB M is dependent on the specified traffic and blocking performance, as well as the order of the shuffle nodes. A typical value of M in the SB is 25 and this is used in the following example in which the substrate areas for electrically and optically interconnected SB networks are compared. Table 2 summarizes these results. The bandwidth per node is assumed to be 250 MHz. As before, the interconnection technology data from Figure 5 are assumed. As with the cross-bar, the interconnection requirements in the metallic technology show a steep growth after N~512, whereas the optically interconnected substrate grows only linearly with N.

Table 2. Comparison of Metallic and Free-space Optical Interconnection Substrate Areas for N node Sliding Banyan Packet Switching Architecture.

Number of nodes, N	FSOI sliding banyan substrate area (cm^2)	Electrically interconnected sliding banyan substrate area (cm^2)
128	.8	0.64
256	1.6	2.56
512	3.2	64
1024	6.4	1024
2048	12.8	4096

In the case of the Sliding Banyan architecture, the freedom of the partitioning of the architecture enables significant algorithmic benefits that add to the substrate area advantages shown in Table 2. These topological benefits provide additional motivation (a factor of ~4 reduction in the number of stages for area-of-interest traffic over other redundant banyan approaches and a great reduction in control resources – by a factor of 1/M) for the optically interconnected architecture [15].

It is important to note that the above interconnection analysis includes only resources connected with the interconnection function within a network. There is, therefore, an implicit assumption that the other resources, such as the size and power consumption of the computing nodes themselves, are dominated by the interconnection resources in measuring performance penalties. The results of this analysis should therefore not be applied without due consideration to the other elements of the network.

The fundamental advantage of FSOI over metallic interconnections in terms of substrate area does not rely on the actual bandwidth densities of the routing layers. It stems directly from the *reduction* in density in metallic interconnections as bandwidth is placed in lower layers. The only technological improvement which would overcome the fundamental advantage is if the lowest routing level (PCB) densities approached the densities of optical interconnections. This is not projected to happen, as density increases tend to "trickle down" from increased chip densities to increased MCM densities, to increased PCB densities. As long as the metallic packaging hierarchy remains the advantage of FSOI will hold true. In other words – although electronic interconnection technology will continue to improve in density (as, we hope, will smart pixel-based FSOI technology), the height and placement of jumps of the metallic interconnect curves depicted in Figure 5 will change somewhat. However the basic and fundamental advantage of FSOI, as embodied in the lower slope and lack of partition boundaries for the optics data of Figure 5, will remain.

This work is sponsored by the Defense Advance Research Projects Agency through a contract with the Air Force Office of Scientific Research.

3.6.4 References

[1] T. Nakahara, S. Matsuo, S. Fukushima, and T. Kurokawa, "Performance comparison between multiple-quantum-well modulator-based and vertical-cavity-surface-emitting laser-based smart pixels," *Applied Optics*, Vol. **35**, No. 5, Feb., 1996.

[2] H. S. Stone, "Parallel Processing with the Perfect Shuffle," *IEEE Trans. on Comp.*, **C-20**, 1971.

[3] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures; Arrays, Trees, Hypercubes*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[4] M. R. Feldman, et. al., "Comparison between electrical and free space optical interconnects for fine grain processor arrays based on interconnect density capabilities," *Applied Optics*, Vol. 28, No. 18, pp. 3820-3829, September 1989.

[5] C. D. Thompson, "Area-Time Complexity for VLSI," in *Proceedings of the 11th Annual ACM Symposium on Theory of Computing*, Atlanta, GA, April, 1979.

[6] A. W. Lohmann, "What Classical Optics Can Do for the Digital Optical Computer," *Applied Optics*, Vol. **25**, pp. 1543-1549, 1986.

[7] G. Eichmann and Y. Li, "Compact Optical Generalized Perfect Shuffle," *Applied Optics*, Vol. 26, pp. 1167-1169, April 1987.

[8] S.-H. Lin, T. F. Krile and J. F. Walkup, "2-D Optical Multistage Interconnection Networks," *Proc. SPIE*, vol. 752, pp.209-216, 1987.

[9] C. W. Stirk, R. A. Athale, and M. W. Haney, “Folded Perfect Shuffle Optical Processor,” *Applied Optics*, Vol. **27**, pp. 202-203, 1988.

[10] A. A. Sawchuk and I. Glaser, “Geometries for Optical Implementations of the Perfect Shuffle,” *Proc. SPIE*, Vol. 963, p. 270, 1988.

[11] M. W. Haney and M. P. Christensen, "Smart Pixel Based Viterbi Decoder," *Optical Computing'95*, March, 1995.

[12] M. W. Haney and M. P. Christensen, “Sliding Banyan Network,” *Journal of Lightwave Technology*, Vol. **14**, No. 5, May, 1996.

[13] R. R. Michael, M. P. Christensen, and M. W. Haney, " Experimental Evaluation of the 3-D Optical Shuffle Interconnection Module of the Sliding Banyan Architecture," accepted for publication in *Journal of Lightwave Technology,* scheduled for September, 1996.

[14] Y. Li, T. Wang, and R. Linke, "A Beam-steering Opto-electronic Cross-bar Interconnect Using VCSEL Arrays," Optical Computing'96, paper OMC2, April 21, 1996PS'96

[15] M. W. Haney and M. P. Christensen, “Sliding Banyan Network Performance Analysis,” submitted to Applied Optics, January, 1996.

3.7 Smart Pixel Algorithmic Tradeoffs for the Sliding Banyan Network

3.7.1.1 Motivation

The Sliding Banyan (SB) has been shown to achieve fundamental performance advantages due to its unique 3D optical topology, which spatially clusters nodes to reduce interconnection and control resources [1]. The SB concept is based on a deflection self-routing scheme, in which a virtual banyan is "slid" to accommodate each packet's routing needs. This clustering approach presents an interesting smart pixel logical design tradeoffs. The tradeoffs discussed in this paper center around the control of the collocated stages of every node in the network. One approach uses a single processor for all stages of a given node. This "node" processor controls all of its stages' switches. Since the SB is a pipelined architecture and the switches are set only once for each packet, the processor needs only control on e stage at a time. A different approach is to place a simpler processor at each stage of the node. There will be many more of these "stage" processors, but they will be responsible only for a single stage's routing, so they can be appropriately simpler. In order to arrive at a simplified design for the stage processor, the required functionality needs to be reduced. In considering these two processor design approaches, the performance of the switching network and the logical complexity required to achieve it are the critical issues which must get examined.

3.7.1.2 Approach

The application we examined was high throughput switching of uniformly distributed traffic. A 1024 node 60 stage SB architecture was modeled [1]. Previous work has shown that this number of stages is more than adequate to successfully route all packets under a variety of stressing traffic patterns, such as area of interest traffic [2]. The OPNET modeling tool was used to model two local smart pixel processing schemes. The basic approach of OPNET is to develop an object oriented model of the switch processor, in this case an array of smart pixel processors. The smart pixel processor is represented as a finite state machine. This representation provides insight into the logical functionality of the smart pixel processors, which themselves are designed as finite state machines.

We simulated two based SB smart pixel control functionalities - each corresponding to a variation of the pipelined routing algorithm, and the nature of the localized control. This first approach involved a node processor which employed a multiplexing strategy to associate all the states of a given node pair tithe the concentrated logic. The 60,000 or so VCSEL/detector pairs, distributed across a large backplane, would e serviced by relatively few (512) smart pixel processing nodes. This node processor implements pure deflection routing with a counter determining packet routing prioritization. The second approach distributed a much simpler processor to each node pair. Therefore, there are 60 times as many processors, but each processor is simpler in design. The simpler processor did not have the full functionality of the concentrated processing unit, so a variation of the routing algorithm was required.

Aspects of the node processor which are costly are the large MUXs on the I/O required to run all stages of a node from a single processor. Also, the deflection routing prioritization requires a counter at each processor. This may prove to be too costly for a fully distributed approach – requiring a counter at every stage of a node in the network. Clearly a different routing prioritization is required; one such scheme would be first-come-first-served. The first packet to arrive at a given stage processor would have priority over the later arriving packet. In this way, the flagging function of the counter in the previous scheme is reduced to a simple time delay on the order of one bit's transmission time. This change in functionality has two implications. The switch may become slightly less efficient since the first packet to arrive is not guaranteed to be the packet closest to its destination – which is a criterion for pure deflection routing. Secondly, the bit synchronous nature of the switching fabric is removed by these randomly occurring time delays, creating a need for a mechanism to determine when a packet is present. Figure 1. is a schematic depiction of the two smart pixel routing control options corresponding to many stage processors or a single node processor.

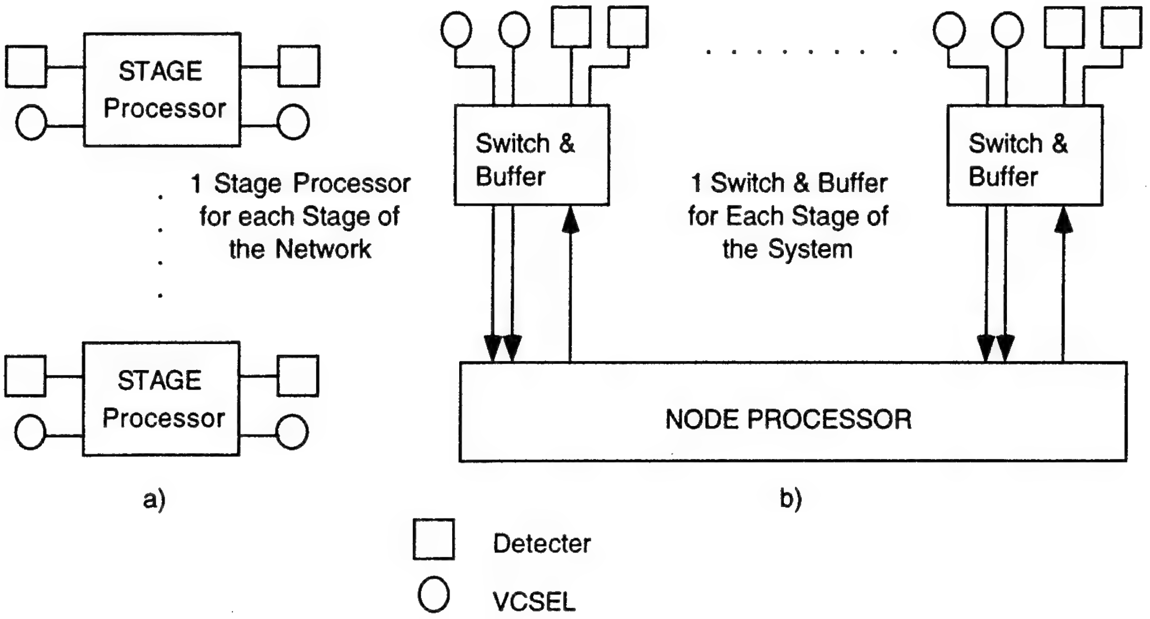


Figure 1. Schematic depiction of two smart pixel SB control options.

Figure 2 shows a typical simulation run of the OPNET model fro the two smart pixel control options. As can be seen, performance is similar with both options routing all traffic in approximately 40 stages. However, the pure deflection routing, as implemented by the node processor, has a slight advantage in routing efficiency. This more efficient routing is a result of the more complicated node processor design. The stage processor was able to approach the performance of the node processor but required additional overhead due to the detection of packets arrival times. Its simpler design and fast decision process led to less overhead at each node and a commensurate reduction in the network delay. The overall latency differed by a factor of ~2 because the complexity of the node processor requires about twice as many clock cycles to route. However, both schemes seem to be within acceptable packet latencies.

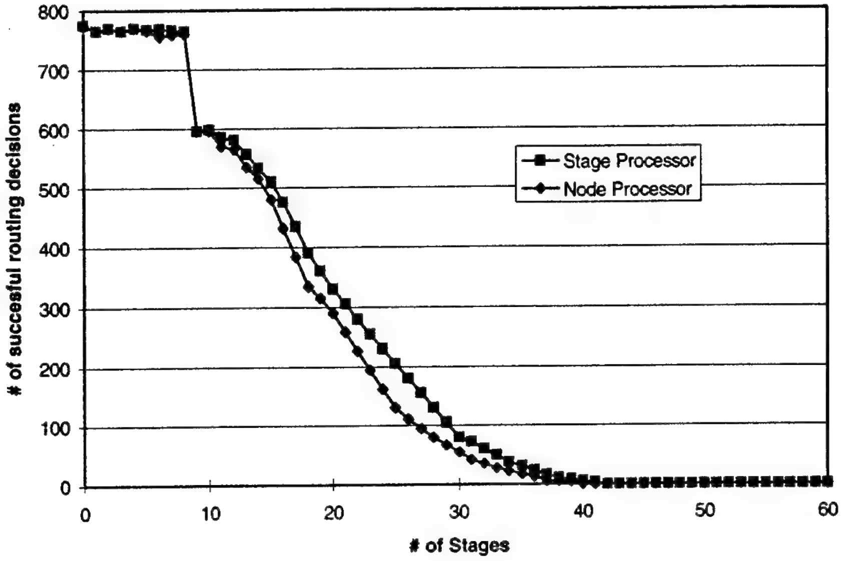


Figure 2. OPNET simulation performance for the two SB smart pixel control options.

3.7.2 Discussion

Given the similar performance of the two control options, the smart pixel complexity must be examined to determine which approach is more beneficial. At each stage of each node both smart pixel designs contain a 2x2 switch with a control line from a clocked D flip-flop, to set and hold the switch until the packet is completely through. The control line is set by the processor during the routing process at that stage. Since both designs require the same number of optical links, the number of VCSELs and detector driver circuits will be the same. However, there is some added complexity in the driver circuit of the stage processor, since additional functionality is required. The main simplification of the stage processor is the elimination of the counter circuit, for prioritization, and the MUXs needed for controlling multiple stages. A complete tradeoff analysis requires gate level resource analysis of the two smart pixel control options, along with further algorithmic refinements through the OPNET simulations.

In conclusion, the OPNET model was used to couple a functional smart pixel state diagram with application performance. This state diagram will be used for VLSI implementations of the SB smart pixel, and will prove useful for the real estate vs. performance tradeoff analyses.

3.7.3 References

[1] M. W. Haney and M. P. Christensen, "Sliding Banyan Network," *Journal of Lightwave Technology*, Vol. **14**, No. 5, May, 1996.

[2] M. W. Haney and M. P. Christensen, "Sliding Banyan Network Performance Analysis," submitted to Applied Optics, January, 1996.

3.8 Two Bounce Free-space Arbitrary Interconnection Architecture

3.8.1 Background / Motivation

Free-space digital optical interconnections have been suggested to overcome interconnection limitations in large, globally interconnected multi-processor architectures [1]. Previous research in this area has focused on the mapping of multiprocessor architectural requirements onto free-space optical interconnection implementations. This has often resulted in a tradeoff between the interconnection an architecture would ideally utilize and those interconnections which the optical implementation readily provided.

The problem of mapping architectures to optical interconnections is worsened when an architecture must perform two or more differing functions. For example, two differing interconnection functions must be implemented to perform a large image 2-D FFT. First, a multi stage butterfly interconnection network must be implemented to perform row and column 1-D FFTs, secondly a corner-turn on the data set must be implemented. The interconnection requirements for these two functions are extremely different. An optical interconnection approach which is designed to perform the first function may not be capable of the second.

A physical optical interconnection approach which provides arbitrary interconnections is desired. As discussed in this paper, without changing lens positions or attributes, a completely different set of interconnections can be achieved – through changing local electronic interconnections. Furthermore, the optical system can be implemented with a symmetric arrangement, thereby allowing the interconnection to be folded back onto itself in a reflective architecture [2]. The details of such an approach are outlined in the next section.

3.8.2 Approach

The two bounce arbitrary interconnection fabric is derived from banyan based multi-stage switching. A banyan is defined as a set of local switching and global interconnection stages, such that every input node has a unique path to every output node [3]. The number of switching stages is $\mathrm{Log}_k N$, where k is the order of the shuffle interconnecting the switching stages. For example, for N=1024 nodes, a banyan comprised of 2-shuffles would require 10 stages ($\mathrm{Log}_2 1024=10$); whereas a banyan comprised of 4-shuffles would required 5 stages ($\mathrm{Log}_4 1024=5$).

If N=k^2 the interconnection fabric is reduced to 2 switching stages interconnected by one free-space optical interconnection. Figure 1 shows banyans based on 2-shuffles and 4-shuffles for N=16 nodes. The 4-shuffle banyan requires only one interconnection stage. Furthermore, it is symmetric, and therefore readily lends itself to an interleaved reflective single plane implementation [2]. Since it is a banyan, every node has access to every other node in a single pass through the optical system. While many architectures may be mapped onto such an interconnection scheme, it is not general enough to provide arbitrary interconnections, or multiple differing interconnections in a single optical system.

Since the interconnection depicted in Figure 1b is a banyan, it suffers from blocking of approximately 70% on average, over all permuted interconnections [2,4]. While every node can reach every other node in the banyan, they cannot do this simultaneously. In fact, the worst case blocking occurs if all nodes on chip 1 want to communicate with all nodes on another arbitrarily selected chip. In this case, all but one of the links are blocked, with (k-1)/k blocking

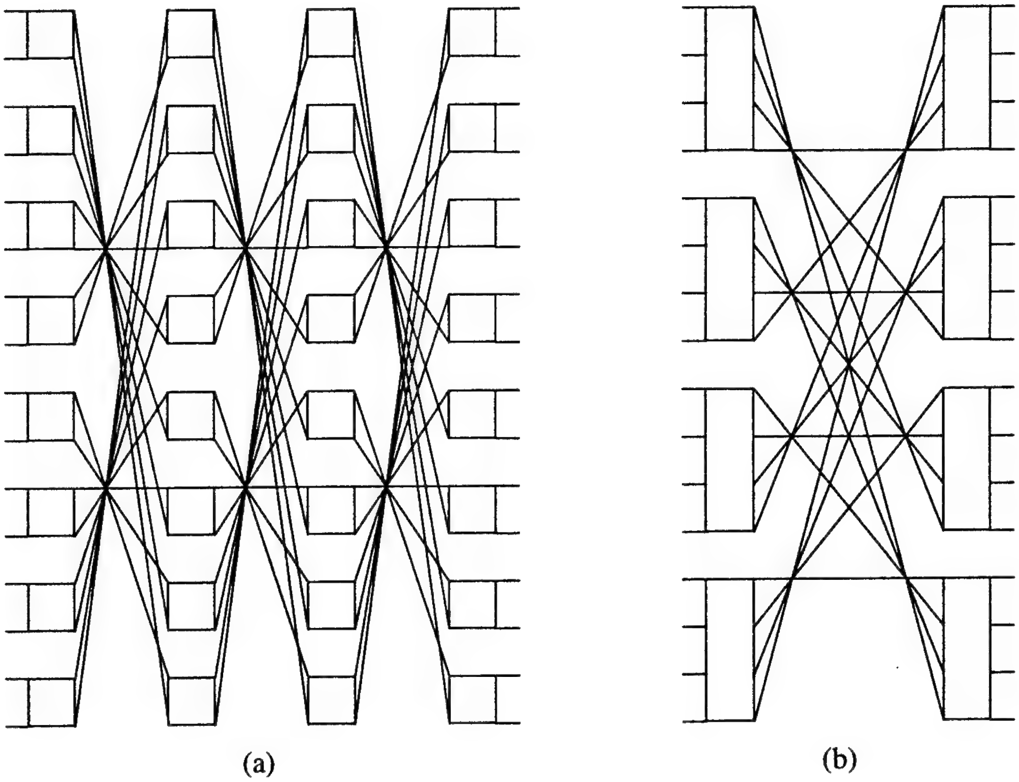


Figure 1a. A banyan comprised of optical 2-shuffles does not lend itself to a reflective architecture due to lack of symmetry. *Figure 1b.* A banyan of k-shuffles (case k=4) is symmetric, and therefore is readily implemented in a reflective architecture

probability.

Banyans are, however, a useful building block for an interconnection fabric. The Benes network [3] is comprised of 2 banyans, placed end to end – one in the forward direction, and one in the reverse. This configuration has been proven to be the minimum rearrangeably nonblocking architecture. Figure 2 depicts the Benes approach when implemented with k-shuffle banyans ($N=k^2$), and 4 OEICs are utilized.

The architecture depicted in Figure 2 can be folded and pipelined, with smart pixel resources contained in a single plane in which the smart pixel I/O are interleaved [4,5]. This topological transformation places the 3 routing blocks depicted horizontally on a single OEIC. There are 4 OEICs in this example implementation. Every input node can communicate to any one output node simultaneously (assuming no two input nodes are attempting to communicate to the same output node i.e., no output contention). This arbitrary interconnection takes 2 passes through the optical system and 3 local electronic routings.

This arbitrary interconnection approach is not useful for direct application to packet switching since the local (on-chip) routing requires global knowledge of the

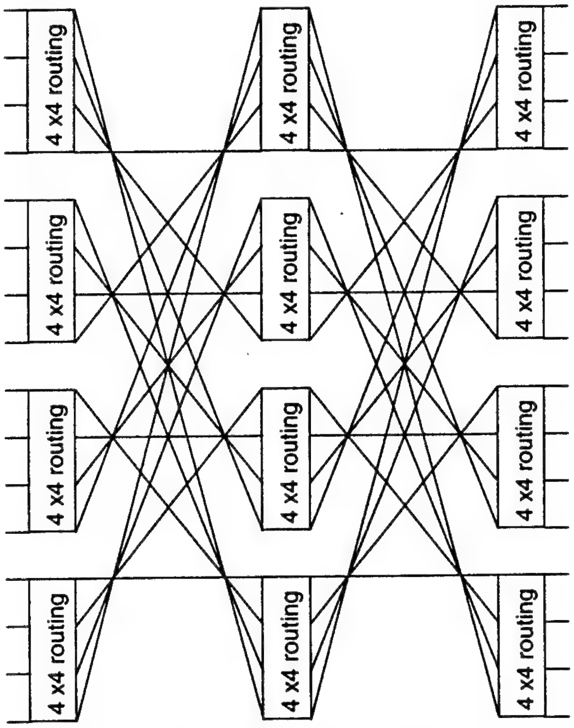


Figure 2. Interconnection fabric is comprised of 3 stages of local switching and 2 passes through a global free-space optical interconnection.

interconnection. However, when the required interconnection pattern is known a priori (e.g., for FFTs, filtering, decoding, and related multiprocessor algorithms), the local electronic routing reduces to simple local (on chip) wiring patterns on the OEICs. If multiple differing interconnection patterns are required, the local routing can be implemented as circuit switching based on look-up tables. While no "live" header decoding is implemented, many differing predetermined interconnection patterns can be paged through.

3.8.3 Discussion

The two bounce interconnection architecture requires simply one extra pass through the optical system to achieve any arbitrary point-to-point interconnectivity among a multiprocessor system. This is the first proposed arbitrary non blocking interconnection architecture based on scalable macro-optics [6]. This arbitrary interconnection comes at the price of a fixed 1 stage latency. The penalty of this extra bounce is significant only when the processor throughput is so high that processors will be waiting for data from the interconnection fabric. Most multiprocessor architectures will require multiple clock cycles for the actual processing of the data. When this is the case, the extra bounce of the optical interconnection is insignificant.

The two bounce arbitrary interconnection architecture relies on symmetric k-shuffle optics and local electronic routing. A multi-chip module (MCM) based retroreflective macro-optical architecture which can implement this has been demonstrated [7] to readily implement the required global k-shuffle interconnection pattern. Such an approach is depicted in Figure 3 for a 4x4 OEIC array with a k=16 shuffle. This interconnection module approach has been shown to scale well for large interconnection bisection bandwidth architectures [6].

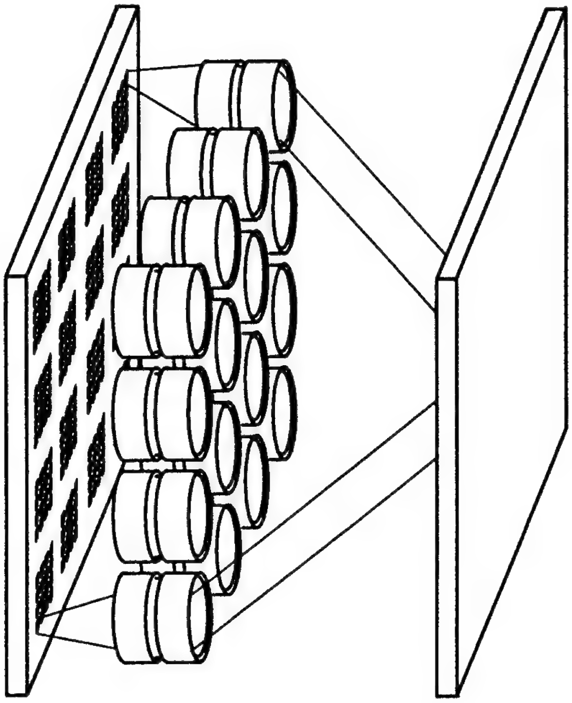


Figure 3. Single plane retro-reflective arbitrary interconnection module.

3.8.4 Conclusion

Through simply adding a single pass through the optical system, a banyan interconnection scheme can be converted into a generalized or higher order Benes – the proven minimum rearrageably nonblocking interconnection fabric. Through applications of topological transformations, this architecture is implemented in a single coplanar MCM stage with the scalability of macro-optical refractive lens elements. This arbitrary interconnection can be fixed upon integration (routing wires) or circuit switched for added flexibility.

3.8.5 References

[1] DARPA/ETO Workshop on Optoelectronic Processing (OEP), May 29, 1996.

[2] M. W. Haney & M. P. Christensen, "Optical freespace sliding tandem banyan architecture for self-routing switching networks," *Digest Int. Conf. Optic. Computing*,. August, 1994.

[3] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes,* 1992.

[4] M. W. Haney & M. P. Christensen, " Sliding Banyan Network,", *Journal of Lightwave Technology*, May, 1996.

[5] M. W. Haney, "Pipelined optoelectronic free-space permutation network," *Optics Letters*, February, 1992.

[6] M. W. Haney & M. P. Christensen, "Fundemental Geometric Advantages of Free-space Optical Interconnections," *MPPOI*, October, 1996.

[7] R. R. Michael, M.W. Haney, & M. P. Christensen, "Experimental Evaluation of the 3-D Optical Shuffle Interconnection Module of the Sliding Banyan Architecture," September, 1996.

4. Conclusions

The FIND program established the viability of FSOI-based solutions to the multiprocessor interconnection bottleneck. The program provided fundamental and significant advancement in the analysis tools, design approaches, and implementation techniques needed to apply smart pixel and FSOI techniques to real-world problems. The significant accomplishments of the FIND program are summarize below.

- Sliding Banyan architecture: The *Sliding Banyan* (SB) network was invented, patented and transfered to a commercialization organization (Capital Photonics Inc.). The concept was validated with 2D and 1D VCSEL arrays. Analytical validation proved dramatic reductions in the number of stages required to acheive a given blocking rate.
- Simulation and analysis tools: The fundemental efficiency of the 3D SB architecture was extended to realistic area-of-interest traffic patterns. Simulations showed the SB approached the provable minimum switching resource requirement and reduced control resources by orders of magnitude. Algorithmic tradeoffs for control of the SB architecture were performed using the telecommunications model OPNET. This allowed a direct comparison of smart pixel complexity without implementation and fabrication of approaches.
- Fundamental scaling laws: The fundemental geometric advantages of FSOI were quantified, for the first time. These advantages were based on projected capabilities of the electronic packaging hierarchy and paractical assumoptions about the abilites of FSOI. The results defined the application domain of FSOI to by the high BSBW domain of multiprocessor architectures. Any problem can be evaluated for its potential gain under and optical implementation by first defining its BSBW. These arguements show that a macro-optical single plane approach is the only one which scales well in size, weight, and power to the multi Terabit regime.
- Interleaved, retroreflective multichip packaging approach: FSOI patterns were shown to be equivalent to topological transformations of graphs. This way of

looking at FSOI provides a framework for exploiting the global interconnection benefits of FSOI. Six basic trnasformations were identified. The SB module combines all six of these to acheive its effieient packaging and performance. One of these transforms was introduced in this program. This is the concept of self-similar grid patterns. Self-similar grid patterns were shown to provide a better match to the geometry of MCMs, thereby overcoming the packaging constraints associated with rectangular grids and providing a ×10 reduction in volume. Also, an arbitrary interconnection approach based on macro-optics was invented. This interconnection fabric allows any multiprocessor architecture to derive advantages from FSOI, without regard to shuffle topology.

- Experimental validation: The SB optical interconnection module was experimentally validated. This prototype acheived 10 um resolution and registration accuracy across a 10 cm MCM like substrate. This was the first experimental demonstration of a retroreflective, multi-chip, single plane, FSOI concept.

Taken together, the accomplishments of the FIND program provide a solid framework for the next step: system level application of FSOI to significant real-world problems in military and commercial arenas.

5. List of Publications

The following are selected journal and conference publications detailing various aspects of the work performed on this project:

[1] M. W. Haney, "Self-Similar Grid Patterns in Free-Space Shuffle/Exchange Networks," Optics Letters, vol. 18, pp. 2047-2049, Dec. 1993.

[2] M. W. Haney and M. P. Christensen, "Sliding Banyan Network," Journal of Lightwave Technology, vol 14, no. 5, pp.703-710, May 1996.

[3] M. W. Haney and M. P. Christensen, "Sliding Banyan Network Performance Analysis," submitted to Applied Optics, Jan. 1996.

[4] R. R. Michael, Marc P. Christensen, and Michael W. Haney, "Experimental Evaluation of the 3-D Optical Shuffle Interconnection Module of the Sliding Banyan Architecture," Journal of Lightwave Technology,vol. 14, no. 9., pp. 1970-1978, Sept. 1996.

[5] Michael W. Haney, "Topological Aspects of Free-space Optical Interconnections," JOP experts workshop, Dec. 1995.

[6] Michael W. Haney and Marc P. Christensen, "Fundemental Geometric Performance Advantages of Free-Space Optical Interconnections," MPPOI 1996.

[7] Christopher R. Osborne, Marc P. Christensen, and Michael W. Haney, "Smart Pixel Algorithmic Tradeoffs for the Sliding Banyan Network," Smart Pixels `96.

[8] Marc P. Christensen and Michael W. Haney, "Two Bounce Free-space Arbitrary Interconnection Network," submitted to IEEE/LEOS Topical Meeting on Optics in Computing, March 1997.